Technical Report 974

Application and Validation of Workload Assessment Techniques

Richard E. Christ
U.S. Army Research Institute

Susan G. Hill and James C. Byers
EG&G Idaho Inc.

Helene M. Iavecchia
Computer Sciences Corporation

Allen L. Zaklad
Chi Systems, Inc.

Alvah C. Bittner
Battelle HARC

March 1993

United States Army Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Acting Director

Research accomplished under contract
for the Department of the Army

Analytics, Inc.

Technical review by

Margaret S. Salter
Joan Dietrich Silver

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1993, March	Final Sep 86 - Sep 92

4. TITLE AND SUBTITLE
Application and Validation of Workload Assessment Techniques

5. FUNDING NUMBERS
MDA903-86-C-0384
62785A
790
1102
C01 & H01

6. AUTHOR(S) Christ, Richard E. (ARI); Hill, Susan G., and Byers, James C. (EG&G Idaho Inc.); Iavecchia, Helene M. (Computer Sciences Corp.); Zaklad, Allen L. (Chi Systems, Inc.); Bittner, Alvah C. (Battelle HARC)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
--

8. PERFORMING ORGANIZATION REPORT NUMBER
--

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-IZA
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARI Technical Report 974

11. SUPPLEMENTARY NOTES
Contracting Officer's Representative, Richard E. Christ.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
--

13. ABSTRACT *(Maximum 200 words)*

This report describes the methods and procedures used, and the findings obtained from a series of eight separate studies across three Army systems; studies conducted as part of an operator workload (OWL) program. It addresses the application of both empirical methods for evaluating the workload associated with the operation of Army systems and analytical methods for predicting that workload. It presents and discusses the results obtained from these studies in terms of their meaningfulness or validity for a number of different practical topic areas.

Two broad conclusions are drawn from the overall OWL program. First, this effort illustrates that it is possible to mount programs to examine research questions in the context of operational and developmental systems. Second, by emphasizing several important workload topics, this report establishes a basis for identifying future research needed for the successful application of workload methodologies.

14. SUBJECT TERMS		15. NUMBER OF PAGES
Operator workload	Human performance	153
Measurement	Workload assessment	16. PRICE CODE
System development	Workload prediction	--

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Technical Report 974

Application and Validation of Workload Assessment Techniques

Richard E. Christ
U.S. Army Research Institute

Susan G. Hill and James C. Byers
EG&G Idaho Inc.

Helene M. Iavecchia
Computer Sciences Corporation

Allen L. Zaklad
Chi Systems, Inc.

Alvah C. Bittner
Battelle HARC

Field Unit at Fort Bliss, Texas
Michael H. Strub, Chief

Training Systems Research Division
Jack H. Hiller, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

March 1993

Army Project Number
2Q162785A790

Human Performance Effectiveness and Simulation

This U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) supports the Army with research and development on manpower, personnel, training, and human performance issues as they affect the development, acquisition, and operational performance of Army systems and the combat readiness and effectiveness of Army units. One concern that underlies all of these issues is the mental workload imposed upon and experienced by the operators of newly emerging, high technology systems and the impact of that workload on operator and system performance. The ARI Fort Bliss Field Unit is conducting exploratory development research to establish the foundation for an operator workload (OWL) assessment program for the Army.

This technical report summarizes the successes and the lessons learned from a series of eight separate field experiments conducted to apply and validate the most promising workload measuring techniques. Because these studies were conducted using three different Army systems, the results that are documented are highly robust with respect to the meaningfulness or validity of the selected workload measurement techniques for a number of different practical topic areas.

EDGAR M. JOHNSON
Acting Director

ACKNOWLEDGMENTS

The operator workload research program summarized in this report was accomplished as a truly cooperative partnership between the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) and numerous individuals who worked at one time or another for the prime contractor, Analytics, Inc., which ceased to exist before final products of the program could be published.

The research program and the preparation of this report benefited immeasurably from the assistance provided by these individuals and others not directly connected with the program. The latter include Sandra G. Hart of the NASA-Ames Research Center and Gary B. Reid of the U.S. Air Force Armstrong Laboratory. Michelle R. Sams, Edwin R. Smootz, Kathryn A. Quinkert, and Julie Hopson of ARI, and Robert J. Wherry, Jr., formally reviewed early drafts of this report; their comments led to major improvements and clarifications. Margaret S. Salter and Joan D. Silver, both of ARI, deserve special thanks for serving as formal peer reviewers for this version of the report; they offered numerous comments that have improved both the content and style of the report.

We owe a large debt of gratitude to members of the operator workload research team who, while not authors of this report, were an integral part of the overall research program. They include Paul M. Linton of Sikorsky Aircraft, John P. Bulger of Science Applications International Corp., Regina M. Harris of Hilton Systems Corp., and Robert J. Lysaght of NYNEX Science and Technology, Inc.

Finally, special recognition is given to Kenneth J. McKeever. Because Analytics, Inc. ceased to exist as an independent entity before the report was completed and because individuals responsible for the research described in this report went their separate ways, the documentation of methods and outcomes of the separate studies were carefully stored in cardboard boxes. The job of unpacking those boxes and making sense of their contents was accomplished through Mr. McKeever's extremely competent efforts. Mr. McKeever, a graduate student at New Mexico State University, was selected to work with technical and analytical personnel in the ARI Field Unit at Fort Bliss, Texas, as a Research Fellow under an innovative program established in 1980 by ARI and the Consortium of Universities of the Washington Metropolitan Area. The Consortium Research Fellows program is an outstanding example of the close and cooperative relationship that can exist between defense laboratories and academic centers; it facilitates and stimulates these types of agencies to combine their research talents and skills to the

ultimate benefit of all concerned. This report might never have been produced without the able assistance of Mr. McKeever and, hence, without the Consortium Research Fellows program.

APPLICATION AND VALIDATION OF WORKLOAD ASSESSMENT TECHNIQUES

EXECUTIVE SUMMARY

Requirement:

In response to the need for useful guidance in the assessment and analysis of operator workload (OWL), the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) sponsored a multiyear exploratory development effort called the OWL program. One objective of the OWL program was to select and apply the most promising OWL measurement techniques to several Army systems. This tech ical report documents the process and outcome of meeting this objective.

Procedure:

A series of eight separate studies was conducted using three different Army systems. These studies applied both empirical methods for evaluating the workload associated with the operation of Army systems and analytical methods for predicting that workload. The empirical methods examined were variants of four operator rating scale techniques. The analytical methods scale techniques and a task analysis and simulation technique. The three systems studied included a mobile air defense missile system, a remotely piloted air vehicle system, and a helicopter system.

Findings:

This report presents and discusses the results obtained in terms of meaningfulness or validity for a number of different practical topic areas. Direct comparisons among the four empirical rating scales showed that one, the Task Load Index (TLX), was consistently highest in factor validity and operator acceptance. For these reasons, TLX is recommended for all but screening applications. The empirical workload ratings are shown to be sensitive to changes in system performance and in the expected levels of workload imposed on the operator by the system, mission, and operational conditions. Additional analyses show that the ratings are robust with respect to delays between a workload experience and its rating and to variations in rater experience with the system under consideration. The TLX subscale ratings are shown to contain potentially useful information concerning the source or cause of experienced workload. Finally, the raw average of TLX subscale ratings is shown to produce composite or

global workload scores essentially equivalent to those obtained using the standard weighted average of TLX subscale ratings.

Both of the analytical methods studied were shown to have promise as methods for identifying potential workload problems early in the system development process. The task analysis and simulation technique was shown to have the capability to track empirical workload ratings. More research is indicated to fully exploit these analytical techniques.

Utilization of Findings:

The findings of these primary data collection efforts added to a foundation of knowledge concerning workload assessment techniques that, in turn, permitted two other objectives of the OWL program to be met. Specifically, these studies contributed to the preparation and publication of two other ARI research products: (a) a computer-based expert system, the Operator Workload Knowledge-based Expert System Tool (OWLKNEST), which provides guidance for selecting the most appropriate techniques to use for assessing operator workload during the systems acquisition process, and (b) a pamphlet for the managers of Army systems that describes the need and some procedures for ensuring that OWL issues and concepts are incorporated into the Army materiel acquisition process. These and other direct outputs from the OWL program have been presented to both scientific and military audiences in over 20 separate papers and symposia at professional meetings and in three edited reference books. Indirect outputs from the OWL program include service as the basis for other programmatic research efforts by such agencies as the U.S. Department of Transportation, as well as literally scores of other related reports and presentations.

Two broad conclusions may be drawn from the overall OWL program. First, the success of this primary data collection effort illustrates that it is possible to mount programs to look at research questions in the context of operational and developmental systems. Second, by emphasizing several important workload topics, this report establishes a basis for identifying future research needed for the successful application of workload methodologies.

APPLICATION AND VALIDATION OF WORKLOAD ASSESSMENT TECHNIQUES

CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

CONTENTS (Continued)

APPLICATION AND VALIDATION OF WORKLOAD ASSESSMENT TECHNIQUES

INTRODUCTION

Purpose of this Report

This report summarizes the information contained in a series of twelve technical memoranda and draft reports. Each of the separate manuscripts describes different studies or phases of a research program that was designed to evaluate the applicability and the validity of operator workload assessment techniques for Army systems. While portions of five of these manuscripts have been previously published in proceedings of annual meetings of the Human Factors Society, they are otherwise unpublished.

There is no attempt in this report to embellish the descriptions and discussions of workload and workload assessment techniques that are given in the previous separate manuscripts. The purpose of this report is to consolidate across the information contained in those manuscripts and to indicate the lessons learned concerning the concept of workload and the methodologies for assessing workload.

Background

The problem. Projected manpower declines coupled with increases in personnel costs and battlefield sophisticati has prompted an increased reliance on high technology equipment in new military systems. As technology has changed, the role of the system operator has also changed. Task requirements for the system operator have shifted from those that primarily require physical exertion to those that demand increasingly larger amounts of perceptual and cognitive exertion.

The relationship between the demands placed on an operator and the operator's capacity to meet those demands constitutes the **workload** imposed upon or experienced by the operator. It has been argued that if the level of operator workload is too great, undesirable, if not catastrophic, consequences may occur. These negative consequences of an **overload** on a system operator might be such outcomes as a risk to soldier safety, a degradation in system performance, or a failure to meet mission requirements.

The concept of operator workload (OWL). The concept of work in the physical sciences is readily understood; work is not performed without some expenditure of energy or other resources, and work rate and efficiency may change depending on the demands of the situation. Likewise for the human, both physical and mental work depend not only on the particular task to be accomplished, but also upon the availability of the internal resources required of the operator to perform the task. Thus, operator workload (OWL) is defined in terms of the interaction between the work imposed on an operator by a task and the operator's capacity to perform that work. (For a discussion of the conceptual foundations of workload see Gopher & Donchin, 1986, and Lysaght et al., 1989.)

The current status of operator workload in the Army. U.S. Army regulations and Department of Defense standards mandate that OWL issues need to be addressed at all stages of the materiel acquisition process. For example, one military specification requires that "... individual and crew workload analyses shall be performed and compared with performance criteria" (U.S. Army, 1979, Section 3.2.1.3.3). The problem with these regulations and requirements is that they provide no systematic guidance to the system developer as to how such a workload analysis should be performed. This lack of guidance has led to the effort that comprises the body of this report. (For a full discussion of military requirements pertaining to workload, see Christ, Bulger, Hill, & Zaklad, 1990, or Hill et al., 1987.)

The operator workload (OWL) program. In response to the need for useful guidance in the assessment and analysis of operator workload, the U.S. Army Research Institute sponsored a three-year exploratory development effort called the OWL Program. The principal goal of the OWL Program was to establish guidance for controlling the workload associated with the operation of Army systems. Its intent was to identify and integrate the most relevant of workload research into a set of practical workload assessment methods for Army systems analysts and managers and then apply and validate these methods on selected Army systems. Lessons learned from OWL studies of these systems would then contribute to the development of guidance on how future workload analyses should be performed.

The OWL Program objectives. There has been considerable research concerned with workload, the majority conducted in laboratory settings. Of the applied research, most has been associated with aviation systems. The challenge of the OWL Program was to apply and validate the most relevant of the workload measurement techniques and use the results to formulate practical guidance. To meet this challenge, five objectives were developed for the OWL Program. These objectives are listed below.

1. Determine the current status of OWL in the Army, including both the formal requirements and the practical needs of Army users.

2. Identify the techniques and methodologies currently available for the assessment of OWL. Analyze the strong points and the disadvantages of each.

3. Select and apply the most promising OWL assessment techniques to several Army systems.

4. Use the results of Objectives 2 and 3 to synthesize guidance as to which OWL techniques should be used for a given system at a given stage in development.

5. Synthesize overall lessons learned from the OWL Program and provide the managers of Army systems what they need to know about OWL.

Research products from the OWL Program. All of the objectives of the OWL Program were successfully met, leading to the publication and distribution of several research products. The more important of these products are given below.

- Hill et al. (1987) presents the results of a review of Army and Defense Department requirements documents and an analysis of interviews with prospective users of the guidance that was to be produced by the OWL Program.

- Lysaght et al. (1989) documents the results of a comprehensive review and evaluation of the concept of workload and methods for its assessment.

- Harris, Hill, Lysaght, and Christ (1992) describes the rationale, capabilities, and features of the Operator Workload Knowledge-based Expert System Tool (OWLKNEST), and gives instructions for using this microcomputer-based tool. The OWLKNEST technology provides guidance for selecting the most appropriate techniques to use for assessing operator workload during the systems acquisition process.

- Christ et al. (1990) is a pamphlet for the managers of Army systems that describes the need and some procedures for ensuring that OWL issues and concepts are incorporated into the Army materiel acquisition process.

As may be seen, these four research products are the outputs of OWL Program Objectives 1, 2, 4, and 5. While numerous briefings and papers were written to present and document the achievements accomplished with respect to Objective 3, the successes and the lessons learned directly from our validation research have not been organized and published as a single research product. The present technical report has been prepared to document the process and outcome of meeting this objective.

Organization of the Report

This report overviews the accomplishments of the original, primary research conducted as part of the OWL Program. It is organized as follows.

- The next section describes the general purpose and the procedures used for the studies that were done. The latter include brief descriptions of the workload assessment techniques used, the three Army systems that served as vehicles for the research effort, and the most salient features of the methods used for each study.

- After the overview of how each study was conducted, the next section summarizes, integrates, and discusses the major results and the lessons learned across all the studies.

- The last section of this report contains the conclusions that evolve from these studies and from the OWL program in general. Included in this section is a discussion of desirable future research in the area of workload.

- More detailed descriptions of the workload assessment methods and the results obtained in each study are included in the appendixes to this report.

OVERVIEW OF THE PURPOSES AND METHODS OF THE OWL PROGRAM STUDIES

The overall plans for the validation and analysis of OWL measurement techniques for selected Army systems are given in Bittner et al., 1987. This section summarizes those plans as they were applied throughout the primary research phase of the OWL program. First, descriptions are given of the general or common purposes and methods of most of the studies. Then, for each of the three selected Army systems, brief descriptions are given for the system and for the purposes and methods which specifically apply to each of the studies conducted for that system.

General Purposes of the OWL Studies

A major purpose of the OWL Program was to evaluate the **applicability** and **validity** of workload assessment techniques for Army systems. The concept of applicability is based upon very practical issues such as how many resources are required to employ a technique and how readily a technique is accepted for use by the proponents and operators of a system. These are matters that may be fairly easily determined.

The concept of validity is a more complex one but equally important. Validation must be examined as a multi-dimensional continuum concerned with the "degree of reality" that can be demonstrated for workload measurement techniques in various situations. That is, how well do the techniques reveal what they are supposed to reveal? In the real world of Army systems, application of a scientific technique can never be fully validated since there are too many uncontrolled variables.

Our approach to validation of a workload assessment technique was to seek and utilize **any and all** information that relates to the "meaningfulness" or operational reality of the OWL technique in question. Such information includes so-called "objective" results, such as how well a soldier or system performs, and so-called "subjective" information, such as a soldier's comments concerning the amount of effort that had to be exerted to perform a task. The goal was to gather all this partial and uncertain information and put it together in a meaningful way.

With this in mind, most of the OWL primary research studies had several purposes in common. In short, whenever the conditions of the study permitted, answers were sought to the following questions.

- What are the relative capabilities and costs associated with the alternative OWL assessment techniques?

- How well do operators accept the administration of the alternative OWL assessment techniques?

- What is the relationship between soldier or system performance and the OWL measures obtained for selected mission segments or tasks?

- Are the OWL measures obtained sensitive to acknowledged differences in workload resulting from crew position and mission segment variables?

General Methods Used for the OWL Studies

There were several common features in the approach used for the primary research studies. These common features include the OWL assessment techniques, the data analysis methods, and the general procedures used to prepare for and to collect the OWL data. A discussion of these three general methodological considerations is presented in succeeding subsections.

OWL Assessment Techniques

A variety of OWL assessment techniques are available and most have been described in previous publications (e.g., Lysaght et al., 1989; O'Donnell & Eggemeier, 1986; Wierwille & Willeges, 1980). As described by Lysaght et al., 1989, these OWL assessment methods may be partitioned into two categories. The **empirical techniques** involve the assessment of workload while the operator is actually operating a simulator, prototype, or representative system, i.e., workload is assessed with the operator-in-the-loop. **Analytical or predictive techniques**, in contrast, may be applied early in the system design process, without an operator in-the-loop. The empirical techniques include those methods which measure the operator's performance, physiological responses, and reports of subjective experiences. The analytical techniques estimate workload through the methods of expert opinion, comparability analysis, task analysis, and simulation.

Empirical techniques. The workload assessment techniques used in the OWL studies were both empirical and analytical. However, only a single type of empirical technique -- operator workload ratings -- was used extensively. As mentioned earlier, these empirical methods are often denoted "subjective techniques" to refer to their presumed weaker reliability compared to other empirical techniques. However, it has been argued that operator ratings are the **most direct** indicators of operator workload (Sheridan, 1980). In this report, this class of techniques is called operator ratings or operator reports.

The other types of empirical workload assessment techniques, primary or secondary task performance measurement techniques and the class of physiological techniques were not used. There are several reasons for this. First, operator ratings are among the most **non-intrusive** of the OWL assessment techniques; they can be administered after the task or mission is complete and hence not disturb the operator during the performance of his or her tasks. Second, operator ratings are very **flexible** and **portable**; no special equipment or data collection devices are needed. Third, operator ratings are **quick and inexpensive** to administer and analyze. Each of these

points is especially important in conducting applied research on fielded systems. In the field, a research effort must fit the usually severe existing constraints -- lack of time and money, last-second changes in important test conditions, lack of experimenter control, and the priority of operational (as opposed to research) needs. Because of these realities, the cited advantages of the operator report methods are very significant.

Based on our research review, we selected four different empirical techniques to use in our studies. They are:

- Task Load Index (TLX) (Hart & Staveland, 1987),
- Subjective Workload Assessment Technique (SWAT) (Reid, Shingledecker, & Eggemeier, 1981),
- Modified Cooper-Harper (MCH) scale (Wierwille & Casali, 1983), and
- Overall Workload (OW) (Vidulich & Tsang, 1987).

Three of these techniques (TLX, SWAT, and MCH) were selected because of previous validation efforts and the OW scale was chosen primarily because of its simplicity. Two of the scales (MCH and OW) are **unidimensional**, i.e., produce only an estimate of overall or global workload. The other two scales (TLX and SWAT) are **multidimensional**, i.e., provide information on the various components or sources of workload, as well as an estimate of global workload. These four scales are each briefly described in succeeding paragraphs. More detailed descriptions and examples of these techniques are given in Appendix A.

The **TLX** obtains ratings of workload on a scale from 0 to 100 (low to high workload) for each of six dimensions: (a) mental demand, (b) physical demand, (c) temporal demand, (d) performance, (e) effort, and (f) frustration. A weighting procedure is used to combine the six individual scale ratings into a global workload score. To account for differences among soldiers in their perception of workload, each operator is required to designate, for each task to be rated, the more relevant dimension of workload from all possible pairs of the six TLX dimensions (a total of 15 pair-wise comparisons). These paired comparisons are obtained prior to the workload ratings. The proportion of times each workload dimension is judged to be more relevant than the other dimensions is used to weight the TLX workload ratings. A unique weighting scale is thus developed and used in the analysis of the TLX workload data for each rater and task to be rated.

The **SWAT** technique obtains ratings on an integer scale from 1 to 3 (low, medium, and high workload) for each of three dimensions: (a) time load, (b) mental effort load, and (c) psychological stress. There are three distinct steps in the use of the SWAT technique. The first, called scale development, requires each operator to sort 27 cards which contain all possible combinations of the three levels of each of the three dimensions. The sort process is designed to produce a rank ordering of the 27 different workload rating outcomes, from lowest to highest perceived workload. Conjoint scaling

procedures are used to develop a single, global rating scale with interval measurement properties based on these clearly ordinal ratings of workload dimensions. The second step, called event scoring, requires the operator to rate the workload of a given task or mission segment using the three SWAT workload dimensions. Finally, in the third step, each three-dimensional rating is converted to a score between 0 and 100 using the interval scale developed in the first step.

The **OW** technique obtains directly a rating of the operator's overall workload experience on a unidimensional scale from 0 to 100 (low to high workload). The unidimensional scale used with the OW technique is essentially the same as any one of the six scales used in the TLX technique.

The **MCH** technique also obtains an overall rating of workload, but less directly than the OW technique. The MCH utilizes a decision tree approach to assist the operator to determine a single, global rating on a ten-point unidimensional scale. The MCH was developed for workload assessment of systems in which the tasks to be performed are primarily cognitive, rather than motor or psychomotor, and for which the original Cooper-Harper scale (Cooper & Harper, 1969) may not be appropriate.

Analytical techniques. Two different analytical techniques -- expert opinion and task analysis/simulation -- were used in two of the OWL primary research studies. As used, they also represented two different approaches to validating analytical workload assessment techniques. The issue of validation is particularly important for the analytical techniques, given the potentially valuable contribution their use can have at early stages in the system design and development processes.

One approach to validating analytical tools, used with the expert opinion technique, is to implement the analytical tool prior to the development of the relevant system or system component, and prior to the execution of the relevant operational or tactical mission of the system. In this approach, the analytical techniques are executed, and, then, when the system ultimately becomes available, the predictions of workload are compared with workload measures obtained using empirical techniques. There are, of course, problems with this approach, not the least of which is matching up the conditions of the empirical test with those that were projected during the analytical phase.

Another approach to validating an analytical technique is to exercise the technique and develop predictions of workload independently of, but simultaneously with, the application of an empirical workload assessment technique. This second approach was used with the task analytic/simulation technique. Here, while the validation effort may be more straight forward, the predictions made will have no great utility or influence since the system (or some facsimile) is already built. The predictions also are made in the context of considerable information about the system -- more than would normally be available during the early system design phase.

Based on our review of workload assessment methodologies, we selected the following two analytical techniques to use in our studies:

- an expert opinion technique based upon the prospective use of the TLX method (Pro-TLX), and

- the task analytic and simulation methods incorporated in the Task Analysis/Workload (TAWL) and the TAWL Operating System Simulation (TOSS) methods (Bierbaum, Fulford, & Hamilton, 1990).

These two analytical techniques are briefly discussed in the next two paragraphs. More detailed descriptions and examples of both techniques are given in later sections and in Appendix A of this report.

The most significant systematic effort to assess expert opinion has been in the prospective application of the SWAT technique (Eggleston & Quinn, 1984; Masline & Biers, 1987; Reid, Shingledecker, Hockenberger, & Quinn, 1984). However, the prospective application of TLX (Pro-TLX) was selected to be used because of previously established superior validity of the TLX assessment technique and because the subjects who were asked to use the prospective technique had much previous training and experience using the baseline TLX technique. Prospective ratings are obtained in a manner similar to their baseline counterparts except the ratings of workload are made in conjunction with descriptions of systems or events that have not yet been personally experienced by the individual making the ratings, rather than systems which the individual has operated in the past.

The TAWL/TOSS technique for predicting workload was selected to be used because, unlike most of the other available task analytic/simulation techniques, it goes beyond a purely time-based definition of workload; it improves the diagnosticity of workload predictions by identifying and predicting workload associated with several behavioral dimensions -- to include cognitive workload demands. The TAWL/TOSS technique has also been successfully used to predict workload for several Army aviation systems -- to include the UH-60A helicopter which is used as a test system for one of the OWL studies.

Common OWL Data Analysis Methods

This section describes salient aspects of the methods used to analyze the OWL data obtained from the OWL Program primary research studies. It summarizes various standard and non-standard statistical data analysis methods and computational analysis software packages used during this phase of the OWL Program, along with a rationale for their use. Several of the non-standard methods are rather novel approaches to addressing specific issues in the program.

Analysis of variance (ANOVA). The ANOVA is used to estimate whether or not certain independent variables and combinations of variables made a significant contribution to the criterion variable (e.g., workload rating). Hence, for example, ANOVA was used to study the effects on workload of: (a) mission variables such as mission segments or tasks, (b) environmental variables such as the presence or absence

of threat activity, and (c) subject variables such as crew or crew position. The ANOVA is used in these cases to estimate the **sensitivity** of workload measures to experimental conditions that varied "known or presumed" levels of imposed workload. The ANOVA has also been applied to provide direct quantitative comparisons of measures. With different measures of workload representing levels of one factor (M) and workload conditions levels of a second (W), the significance (and follow-up analyses) of the MxW interaction in principle provides a direct comparison of the sensitivity of the different measures. This latter use of the ANOVA requires both that the measures be statistically commensurable, and that statistical adjustments be made (see Bittner et al., 1987, p. 9).

Correlation and regression analysis. These data analysis methods are a useful alternative or follow-up to the ANOVA. The ANOVA determines whether or not a given independent variable contributes significantly to variations in a dependent or criterion variable. On the other hand, correlation methods provide estimates of the degree of relationship between any two variables and regression methods compute the best linear relationship (i.e., the best-fitting straight line) between any two variables. In multivariate analyses there are more than two variables (more than two measures or scores for each subject). Regression analysis was used in the OWL studies, when possible, to determine the relationship between measures of workload and measures of performance.

Factor analysis. Factor analysis methods, and, more specifically, principal components analysis (PCA) represent a class of statistical techniques, based on correlations, which determine the underlying structure of a set of data. In particular, factor analysis computes the "dimensionality" of a set of data (i.e., a minimal set of underlying factors); in practice, these factors are related to meaningful psychological concepts, if possible.

In all of the OWL Program studies reported here, factor analysis revealed a single factor underlying each of the various sets of workload data. This common factor -- the **"OWL Factor"** -- is the result of a linear combination of the standard unit scores from each set of ratings. It is often used to evaluate the effects of workload in the OWL studies, rather than the operator ratings obtained by using any specific rating technique, since it represents the best possible estimate of whatever is being measured by the rating scales.

As a principal method for directly comparing the alternative workload assessment techniques, the OWL factor was correlated with the workload ratings obtained with each of the different types of operator rating techniques. The correlation of each technique's rating data with the OWL Factor is the **Factor Loading** or **Factor Validity** of a particular technique. The factor loadings are measures of the sensitivity of the various techniques in this situation.

Jackknife methods. The Jackknife methods (see Hinkley, 1983) are techniques for closely examining individual differences in conjunction with standard analyses such as ANOVA. Using the Jackknife, the data from each subject are removed (with replacement) one-by-one from the data set and the ANOVA (or other technique) is

applied to the remaining data. This results in N analyses (for N subjects), each of which is missing the data from a different subject, thus assessing the relative contribution of each subject. In the OWL studies which used multiple types of operator ratings techniques, Jackknife methods were used with factor analysis to evaluate the effects of individual operators on the resulting factor loadings of the techniques. This Jackknife analysis provides a measure of the stability of the estimates of the factor loadings in the form of a loadings (one per technique employed) by subject dropped matrix which could be analyzed by a conventional repeated measures ANOVA to determine if there were any significant differences among the factor loadings.

Statistical software packages. For relatively large sets of data or sophisticated analyses, computerized statistical analysis packages are used. For the OWL program studies, BMDP Statistical Software (1987 Release for the Zenith personal computer: Dixon, 1983) was used. The BMDP2V program was used for ANOVA, BDMP2R for regression, and BMDP4M for principal components analysis.

Common Procedures for the OWL Studies

The real-world settings of the OWL studies required careful planning and coordination with the proponents of the system, the operators who were to participate in the test, and various field authorities (e.g., the test officer). For each study, the OWL data collection team became as knowledgeable as necessary and possible about the system and its operational environment. The need to adequately prepare for a study often required multiple trips to the field site as well as the conduct of multiple pilot tests prior to a data collection effort.

Prior to the start of these data collection efforts, an initial briefing and orientation session, lasting a minimum of two hours, was conducted with participating soldiers. These meetings had several purposes: (a) to introduce the OWL team members and legitimize their participation in the data collection effort, (b) to define workload and give instructions and training on the use of the workload rating scales that were to be used, and (c) to obtain demographic and other data, to include, as appropriate, SWAT card sort or TLX paired-comparison data, for use in later analyses.

The OWL data collection effort was almost always an adjunct to a field test or exercise. Therefore, it had to be planned and executed in a manner that would not interfere with the primary activity of the soldiers. The physical and emotional states of the subjects also were taken into account by the OWL data collection team. Since the soldiers were available to the OWL team only after they had just performed a long, hard mission, the data collection environment was designed to provide them with a sense of rest and relaxation (e.g., "soft drinks and chips" were generally made available). The participating soldiers were also isolated as much as possible from other test personnel to protect them from those who might wish to attribute problems in system or unit performance to "subject error," rather than to the design of the system or test.

Studies Using the Forward Area Air Defense (FAAD) Line-of-Sight-Forward-Heavy (LOS-F-H) System

Five individual OWL investigations were conducted on the FAAD prototype LOS-F-H system during and between two field tests in 1987 and 1988. The first study was a retrospective assessment of OWL conducted 10-weeks after a field test which was part of a non-developmental item candidate evaluation (NDICE) system procurement program. This first study is reported by Hill, Zaklad, Bittner, Byers, and Christ (1988). The second study was designed as a follow on to the first and addresses the OWL associated with generic missions; it is reported by Bittner, Byers, Hill, Zaklad, and Christ (1989). The third and fourth studies were based on two different segments of a Force Development Test and Experimentation (FDTE) program. The third study assessed OWL at the conclusion of each of a series of 4-hour missions, and was reported by Hill, Byers, Zaklad, and Christ (1989b); the fourth, at the conclusion of two different 48-hour missions, is reported by Hill, Byers, Zaklad, and Christ (1989a). The fifth and final data collection effort was a prospective assessment of OWL in which operators were asked to predict the workload they would experience with potential improved versions of the system or with revised organizational and operational configurations of the system. The last study was reported by Hill, Byers, Zaklad, Bittner, and Christ (1988). A separate paper was prepared by Byers and Hill (1989) to describe the comparison of individual workload ratings of crew members and the field test performance of the LOS-F-H system during the FDTE. Another report which described all five of these studies was prepared by Hill, Byers, and Zaklad (1989).

LOS-F-H System Description

The LOS-F-H component of the FAAD system will provide air defense support to maneuver elements of a close combat combined arms division. The LOS-F-H system must provide a full range of air defense capability in meeting the low-altitude helicopter and fixed-wing air threat which ground maneuver elements face, and must have mobility and survivability equivalent to the type of force being supported. The baseline or prototype LOS-F-H was selected from among four off-the-shelf (i.e., non-developmental item) candidates provided by various teams of contractors. This pre-production model of the LOS-F-H became the focus of five OWL studies described in this report.

The prototype LOS-F-H was mounted on a M113 armored personnel carrier. It had detection and tracking capabilities consisting of radar, two electro-optical sensors (TV and FLIR), a laser range finder, a laser for missile guidance, and associated consoles for a commander/radar operator (RO) and a gunner/electro-optics operator (EO). The system is operated by a crew of three soldiers who have the following, respective, responsibilities:

- Radar Operator (RO): commands the fire unit and crew, supervises all crew functions and tasks, and performs critical tasks during target engagement sequence, to include those associated with target detection, identification and

prioritization, and the hand-off of the target to be engaged to the EO;

- Electro-optics Operator (EO): assists in target detection and identification, acquires and tracks the target, and fires missiles; and

- Driver (DR): drives the LOS-F-H vehicle during movements and assists in target detection and the selection of movement routes and battle position.

LOS-F-H NDICE Study

The major purposes of the LOS-F-H NDICE study were to directly compare alternative workload rating techniques and to evaluate the relationship between system performance and the "retrospective" workload ratings of the crew members for specified mission segments and tasks.

Workload assessments using each of the principal rating techniques (i.e., TLX, SWAT, MCH, and OW) were provided by six operators of the baseline LOS-F-H system, **10 weeks after their participation in a field test** conducted to support the NDICE. Five operators were EOs and one was a RO during the NDICE field test. The workload assessments were made in conjunction with a review of videotape (with sound) recordings of the two particular mission vignettes selected for evaluation. Time-locked video monitors provided independent views of the RO and EO primary display and control consoles.

Across the two mission vignettes, operators rated the workload they experienced during an attack by two fixed-wing aircraft, two rotary-wing aircraft, and one rotary-wing aircraft. Within each attack sequence, workload ratings were made of three task segments: detect/visual identification, target handoff from the RO to the EO, and target tracking. In addition, overall workload ratings were obtained for both mission vignettes and for the entire NDICE. System performance scores were 0, 1, or 2, reflecting the number of targets successfully engaged during a given attack sequence. Detailed descriptions of the methods and the results of this study are given in Appendix B.

LOS-F-H Generic Mission Study

The previous LOS-F-H NDICE study focused on obtaining estimates of workload made after watching videos of the operator's own performance. Though performance and workload were related, workload ratings were not affected by mission conditions (e.g., type of attacking aircraft). Rather, it appeared that the ratings reflected idiosyncratic differences in the specific mission conditions. The resulting variation in workload experiences washed out the effects of mission variables. The approach taken to overcome such mission-specific "quirks" (and the small number of data points) was to collect workload ratings of generic or "average" missions. This study also explored the difference in workload ratings between operators (LOS-F-H crew members) and other kinds of subject matter experts (SMEs).

Two groups participated in the Generic Mission study: five system operators (EOs only) and nine SMEs. The SMEs were civil service and contractor civilians who had been or would be working directly in the LOS-F-H program. Mission conditions considered were: (a) a single rotary-wing attack, (b) dual rotary-wing attack, and (c) a dual fixed-wing attack. Tasks for each type of mission were (a) visual target identification, (b) target handoff, and (c) track-to-intercept. The nine combinations of mission conditions and tasks were defined, and the subjects were asked to rate the workload associated with each combination, using each of the four rating scales selected for evaluation in the OWL Program studies (i.e., TLX, SWAT, MCH, and OW). These ratings were to be based on the rater's total experience with the system during a previous field test; SMEs not familiar with the LOS-F-H field test were asked to base their ratings on their knowledge of similar systems and tests. The crew members (all EOs) made workload ratings only for specific tasks which they actually performed; SMEs made ratings for prescribed RO and EO tasks. Detailed descriptions of the methods and the results of this study are given in Appendix C.

LOS-F-H FDTE Basic Study

The purpose of the FDTE field test was to examine tactics, doctrine, organization, and training that had been developed for the LOS-F-H system. The test took place over a six-week period. The first five weeks were composed of four-hour missions and the last week was devoted to two 48-hour missions. The FDTE "basic" study investigated workload ratings during the four-hour missions.

The system operators were organized into two crews, with one RO and two EOs in one crew and the other RO and three EOs in the other. The ROs operated solely in that duty position while the other crew members rotated between the EO and driver positions. These seven operators had participated in the previous field test of the LOS-F-H (i.e., the NDICE test) and had served in previous workload studies in the OWL Program.

The field test investigated the performance of crews for mission segments that were documented in battle drills. The mission segments tested were: (a) prepare for road march, (b) road march, (c) emplacement, (d) target acquisition/tracking, (e) reload, and (f) one-man acquisition/tracking operations. Several operational or environmental variables of interest (e.g., day and night missions) were systematically changed over the duration of the test. Upon completion of a four-hour mission, the crew members were taken back to a debriefing room at the base camp where the workload data for the mission just completed were collected. During the first two weeks of the FDTE Basic study, workload ratings were made using each of the four scales selected for evaluation in the OWL Program studies (i.e., TLX, SWAT, MCH, and OW); during the final three weeks, ratings were made using only the TLX and OW techniques. Detailed descriptions of the methods and the results of this study are given in Appendix D.

LOS-H-F FDTE 48-Hour Mission Study

Following five weeks of four-hour missions, the FDTE examined performance in 48-hour missions designed to emulate the operational mode summary for the LOS-F-H. Two three-man crews participated, one crew in each of the two extended duration missions. The two different 48-hour missions were conducted at different times. However, the schedule of events planned for both missions was the same, and included 14 road march, eight acquisition/tracking, and six reload mission segments. With only minor exceptions, all events took place approximately as scheduled.

At periodic times during the 48 hours, the crew was asked to give OWL ratings using only the TLX and OW rating scales. The OWL measures asked for a rating of the workload of the "Overall Mission So Far." Two formal debriefs of the crew took place during the mission. The first took place in the field after the first 24 hours, the second after the completion of the 48-hour mission. The debriefs provided an opportunity to gather additional OWL ratings on engagement-specific tasks and more general conditions. Detailed descriptions of the methods and the results of this study are in Appendix E.

LOS-F-H Prospective Study

The first four studies conducted for the LOS-F-H baseline system were concerned with application of **empirical workload assessment techniques.** These techniques permit measurement of the workload experienced by crew members when they operate a system that has already been at least partially developed. Of equal or greater importance, are the **analytical techniques** which may be used at the earliest stages of the system design process. The analytical techniques may predict operator workload experiences in systems or system applications that have not yet been developed or exercised with an operator-in-the-loop. One analytical technique that needs further investigation is called prospective or projective workload ratings.

Prospective OWL ratings were obtained using only the NASA TLX scales at the conclusion of FDTE. They were obtained using the six soldiers who had been LOS-F-H operators during both the NDICE and FDTE tests. In conjunction with descriptions of systems or events that have not yet been personally experienced by an operator, these prospective ratings were used to predict workload for several critical issues in LOS-F-H system development. It was anticipated that these predictions would be empirically validated in later field tests.

Four distinct topic areas were chosen for prospective investigation. These were (a) new radar equipment which would automate many tasks currently being performed manually by the RO, (b) multiple LOS-F-H fire units, (c) instances of multiple threat targets appearing in rapid succession, and (d) new crew organization. New equipment and crew organization represent optional system modifications, whereas multiple fire units and multiple targets reflect a more realistic tactical context.

The prospective workload ratings were obtained during the sixth and seventh weeks of the LOS-F-H FDTE. While one crew was participating in its 48-hour mission, the other was performing prospective ratings. In turn, each prospective topic area was described and its workload estimates obtained. Detailed descriptions of the methods and the results of this study are given in Appendix F.

Studies Using the Aquila Remotely Piloted Vehicle (RPV)

Two separate workload studies were conducted for the Aquila Remotely Piloted Vehicle (RPV). The first study was conducted during a Force Development Test and Evaluation (FDTE) and is reported by Byers, Bittner, Hill, Zaklad, and Christ (1988). The second study was conducted in conjunction with a tactical deployment of the Aquila RPV and was originally reported by Byers, Christ, Hill, and Zaklad (1988). A separate report which described both of these Aquila studies was prepared by Byers, Hill, Zaklad, and Christ (1989). This section is based on the information contained in these earlier reports.

Aquila RPV System Description

The Aquila system was a remotely controlled air vehicle and payload system designed to be an eye in the sky for field commanders. It could provide field commanders with real-time reconnaissance and surveillance information at ranges five kilometers beyond the forward line of friendly troops. Specific designated functions of the Aquila system included target acquisition, target designation, artillery adjustment, post-mission fire assessment, and intelligent battlefield management. (A detailed description of the Aquila RPV mission, system, and organizational and operations plan, to include diagrams and drawings, is given in Bittner et al., 1987. What follows is a condensation of that more complete description.)

The major components of the Aquila system were: the remotely piloted air vehicle (AV); the mission payload (MP) carried by the AV, which included camera, communication, and designation equipment; a hydraulic launch system which propels the AV to flight speed; a recovery system comprised of a dacron net into which the AV is flown at the end of its flight; the ground control station (GCS) in which the equipment items and personnel necessary to operate the AV and MP were located; and the remote ground terminal (RGT) connected to the GCS by a fiber optics cable. The RGT transmits information between the GCS and the Aquila RPV while the latter is in flight.

The GCS is the operations and control center for the Aquila system. It contains three duty positions. Individuals assigned to these positions perform critical tasks that determine the success of Aquila operations. The initiation of an Aquila mission begins when the crew of the GCS receives a mission order. A key element in the receipt of mission orders included an evaluation of those orders by the GCS crew to identify and resolve conflicts such as incomplete orders, high-risk missions, and inexecutable missions. After the mission orders have been received, the GCS crew must develop detailed

mission plans (to include AV launch and recovery, flight profile, and camera parameters), and manually compute and estimate critical time and location parameters (to include anticipated hovering and search strategies). The detailed mission plans must be entered into the GCS main computer system along with site survey and weather data. The mission planning activities must be successfully and expeditiously completed to meet a requirement that the AV be launched within one hour of receiving the mission.

Shortly after the AV is launched, its control is handed off to the air vehicle operator (AVO) in the GCS, who must continuously monitor its status and position, and maintain linkage to the AV through the RGT. When the AV is positioned over a target area, crew members must perform several operations. These include detecting, recognizing, and locating targets of military significance (a set of tasks principally under the control of the mission payload operator [MPO]), and communicating target information to units requiring it (a responsibility of the mission commander [MC]). In addition, the MPO and MC, in particular, may be required to designate targets for precision guided munitions, to call for and adjust fires, and assess damage to targets which have been engaged. These RPV functions may be required for each of several areas of interest during a single mission. As the RPV mission draws to an end, the AVO directs the flight of the AV toward the location of the recovery net, and an automatic system in the recovery system controls the final recovery of the AV.

The Aquila system was in development for over 10 years. Three events during that development were relevant to the OWL Program. A brief description of the methods and results of the first event (i.e., an operational test) is given below since is serves as background for the second event. The last second and third events were occasions for the conduct of workload studies during the OWL Program. A summary the purpose and methods of those two studies is presented after the description of the operational test.

Aquila Operational Test II (OT II)

The Aquila OT II was conducted from November, 1986, to March, 1987, at Fort Hood, Texas. The OT II was a major evaluation of the Aquila system. It involved 138 missions in which the AV was launched, flown over a battlefield area to perform all of its designated functions, and subsequently recovered. During the OT II, the GCS crew consisted of three soldiers, the AVO, the MPO, and the MC. The AVO and MPO held ranks of Private First Class through Sergeant; the MC was a senior Non-Commissioned Officer or Warrant Officer.

The preliminary results of the Aquila OT II suggested a very low target detection rate. Many factors may have contributed to this result, one of which was that the crew had a difficult time searching an entire designated area. One cause of this problem was that if the crew caused the AV to depart from a planned search pattern to further investigate targets or target areas of interest, they could not easily return to the point in the search pattern where they had left off. Furthermore, there was evidence that the GCS crew did not appropriately interact with representatives of the higher echelon

command group to determine which aspects of a proposed Aquila mission were within and which were outside the system's capabilities. For example, the designated search a : was sometimes larger than could be accommodated by the capabilities of the Aquila system. These preliminary findings implied that the system and its operational procedures had serious problems that should be further investigated and resolved before a production decision was made. While the Aquila OT II was conducted prior to the start of the OWL Program, there was an opportunity to assess workload experiences of the GCS crews during a subsequent test of the system, as described in the next section.

Aquila FDTE Study

The preliminary results of the Aquila OT II established a need for an Aquila FDTE. That earlier test suggested that the GCS crew members could not adequately detect, recognize, and locate targets. Accordingly, the FDTE focused on the ability of the GCS crew to plan and execute an RPV search mission. In addition to providing special training to the crew members, new hardware and software were added to the GCS computer to assist in the process of planning and searching for targets. Also, principally to improve the crew's ability to plan missions, a fourth member was added to the crew. A commissioned officer (i.e., a lieutenant) was assigned to the position of mission commander. The senior non-commissioned officer or warrant officer who had filled that role was named to a loosely defined position of RPV technician. There was no change in the personnel assigned to the AVO and MPO positions. Finally, to reduce the risks associated with launching, flying, and recovering the RPV, the mission payload package was mounted to the underside of a highly maneuverable aircraft; the pilot of the manned aircraft responded to signals that would normally have been sent to the air vehicle.

Operator workload ratings were obtained from 17 GCS crew members, four complete crews and one replacement soldier. Each crew member made individual ratings of OWL during post-mission sessions for each of the five or more missions which were planned and flown by his crew. Two segments of each mission were always rated: Mission Planning and Flight. The four workload rating scales selected for evaluation in the OWL Program studies (i.e., TLX, SWAT, MCH, and OW) were administered in counter balanced order over successive missions, crews, and crew members. Detailed descriptions of the methods and the results of this study are given in Appendix G.

Aquila FIREX 88 Study

FIREX 88 was a major live-fire artillery exercise held in June, 1988, at Dugway Proving Ground, Utah. During its employment in FIREX 88, Aquila was used tactically, for the first time in its history, rather than in a test and evaluation context. The tactical objectives of the Aquila system during FIREX 88 were to perform target detection, recognition, and location, call for fire, and fire spotting tasks. In addition, an ancillary objective of the Aquila battery was to introduce and demonstrate the capabilities of the RPV to senior military commanders and other interested parties.

Workload was assessed using only two rating scales (TLX and OW). Workload ratings were obtained for 15 GCS crew members, three Remote Ground Terminal (RGT) crew members, and three launch/recovery system crew members. With overlap of crew members, a total of 19 subjects provided workload ratings. Each GCS crew consisted of its customary three members (i.e., the MC, AVO, and MPO, as given above in the Aquila system description). During FIREX 88, however, there were as many as five crew members working in the GCS, as training in all three duty positions was ongoing.

Individual workload ratings were obtained from the GCS crew immediately after the conclusion of each of seven Aquila missions spread out over four days. Each of the seven missions had a different crew configuration. Three or four mission segments were rated for each mission; they were Launch, Flight Operations, Recovery, and, when appropriate, the Flight Operation sub-segment of Target Location/Call for Fire.

Individual workload assessments for the RGT and for the launch and recovery systems were obtained near the end of FIREX 88. Three individuals rated RGT workload for two segments: Power-up and Align. Another three individuals rated launch/ recovery workload for four segments: Activate and Checkout the Launch Subsystem, Conduct Launch, Activate and Checkout the Recovery System, and Conduct Recovery. The workload assessments for the RGT and launch/recovery systems did not reflect workload on any one mission but rather an average workload over all the FIREX 88 missions. Detailed descriptions of the methods and the results of this study are given in Appendix H.

Studies Using the UH-60A Black Hawk Helicopter 2B38 Flight Simulator

One two-part study was performed for the UH-60A Black Hawk system. During one part of the study empirical measures of OWL (i.e., operator workload ratings) were obtained from crew members during and immediately after each of two one-hour missions in the UH-60A 2B38 flight simulator. During a second part of the study, an analytical model of the UH-60A was updated and then executed for a mission that matched that used during the empirical data collection runs on the flight simulator; the predictions of the model were compared to the operator ratings. This study was documented in an unpublished technical report (Iavecchia, Linton, Harris, Zaklad, & Byers, 1989). A paper which compared empirical operator workload ratings with predictions of the analytical model was reported by Iavecchia, Linton, Bittner, and Byers, 1989).

UH-60A System Description

The U.S. Army's UH-60A Black Hawk is a twin-engine rotary-wing utility helicopter designed specifically for combat and combat support missions comprised of tactical transport of soldiers, troop units, and required supplies and equipment. Cockpit,

instrument panels, and interior lighting are all designed to accommodate both day and night full-mission capability. The flight control system provides maneuverability for low level, nap-of-the-earth flying. The basic UH-60A crew consists of a pilot, copilot, and crew chief/gunner. The aircraft has virtually identical control and display configurations on either side of the tandem cockpit, and can be properly flown by either the pilot or copilot.

The UH-60A 2B38 flight simulator consists of a molded two-piece cockpit mounted upon a large motion platform. The front cockpit is a faithful reproduction of the fielded UH-60A unit consisting of a pilot and copilot station; behind the flight stations is an instructor/operator station, and an observer station. The cockpit assembly is mounted upon a motion system which provides dynamic movement and accurate cues for pitch, roll, and yaw, along the vertical, lateral, and longitudinal axes, as well as any combination thereof. Four out-the-window cathode ray tube-based displays are provided for the pilot and copilot stations. The displays allow forward and side viewing of a simulated environment during dawn, day, dusk, night, and night vision goggle (NVG) conditions.

OWL Measures

Five operator workload rating scales were used: the four workload rating scales selected for evaluation in other OWL Program studies (i.e., TLX, SWAT, MCH, and OW), and a scale developed specifically for this study, peak workload (PW), modelled after the OW scale. The PW scale was constructed to tap the operator's momentary experience of the highest level of workload over the duration of a mission segment or task.

The analytical model chosen to make predictions of workload was based on the TAWL/TOSS technique. This analytical tool requires inputs which include: (a) a detailed task analysis defining the low-level task activities required for each mission-essential task (e.g., control altitude) together with the task times; (b) estimates of the level of workload in each of five information processing channels (i.e, auditory, visual, kinesthetic, cognitive, and psychomotor) for each low-level task on a scale from 0 to 7 (very low to very high workload); and (c) a set of scenario decision rules to drive the tasks to be performed during each half-second simulation time interval, to include the probability of random concurrent tasks. Given these inputs and the generated time line of low-level task activities, TAWL/TOSS adds the workload values within each channel for concurrent tasks. If the sum of channel workload values across tasks for any half-second interval exceeds a value of 7, an overload is defined to have occurred for that channel.

Procedure for Simulator Data Collection

Seven two-man crews successfully provided empirical OWL measures. All subjects were experienced UH-60A aviators and were currently assigned as instructor

pilots (IPs) at the U.S. Army Aviation Center. Two additional senior IPs were selected to rate the performance of the pilot and copilot during the simulator trials and to assist in the collection of real-time workload ratings. Each crew flew two experimental flights -- one day mission and one NVG mission. The two missions were essentially the same although the night mission was confined to a smaller, as well as different, geographical area to accommodate the slower speeds flown at night.

During these simulated experimental flights, the primary task of the pilot was limited to flight management and of the copilot, navigation and communications. Once a mission was underway, the controller IP asked both operators to report in near real-time the OW and PW experienced during each of twelve mission segments. The controller IP also rated the performance of both operators for each segment. Following each experimental flight, the two crew members gave retrospective workload ratings for all twelve mission segments using the OW and PW scales and for only four selected mission segments using the TLX, SWAT, and MCH techniques. Following the post-mission period of rating workload, a structured interview was conducted with both crew members to assess operator acceptance of the various rating techniques and to gather other general comments.

Procedure for TAWL/TOSS Data Collection

The required updating of the baseline TAWL UH-60A model was independently accomplished by personnel from Anacapa Science, Inc. (D. B. Hamilton & C. R. Bierbaum, personal communication, December, 1989). Specifically, the model had to be modified so that the operator tasks and decision rules would reflect the specific mission requirements of the simulated experimental flight. Only the day mission parameters were incorporated into and executed by the TAWL/TOSS model. Seven iterations of the TAWL/TOSS model were executed. The average output of all runs was used to generate TAWL/TOSS-derived OW and PW measures. To derive a TAWL/TOSS-based estimate of OW for each mission segment, the TAWL/TOSS workload values for each half-second interval within a mission segment were averaged over all five TAWL/TOSS channels. The derived (or predicted) OW score was the mean of these half-second values over the duration of the mission segment. To derive a TAWL/TOSS-based estimate of PW for each mission segment, the TAWL/TOSS workload values for each half-second interval were summed across the five TAWL/TOSS channels. The maximum value of all half-second summed values was defined as the PW for that segment. More detailed descriptions of the methods and the details of the results of this study are given in Appendix I.

RESULTS AND DISCUSSION

This section gives summary descriptions of a number of results obtained from the OWL Program primary research on workload assessment techniques. The emphasis will be on the results which relate to the measurement techniques themselves. Results which are unique to the three test systems* will be reported here only in so far as they demonstrate the viability of the workload measures and the workload measurement techniques. The results obtained for the empirical techniques are summarized first, followed by those obtained for the analytical techniques.

Direct Comparison of Empirical Workload Assessment Techniques

Four operator rating techniques -- TLX, SWAT, MCH, OW -- were directly compared with each other along several dimensions:

- Factor validity,
- Operator acceptance,
- Resource requirements, and
- Special procedures.

Succeeding sub-sections will present the findings obtained for each of these types of comparisons.

Factor Validity

The analysis of factor validity was conducted in two stages. During the first stage, factor analysis was performed on the aggregated data from each study to examine how each of the four rating scale techniques was able to discriminate among different levels of task loading. More specifically, principal component analysis (PCA) was conducted on all possible sets of workload measures collected across all subjects, missions, mission segments, and tasks within each study. Each set of workload measures included global workload rating values derived from using four scales: TLX, SWAT, OW, and MCH. The BMDP4M program (Dixon, 1983) was used to perform these analyses. The results of these analyses are shown in Table 1. Across all the studies shown in this table, the factor analyses revealed a single component variable, hereafter termed the **OWL Factor**, which explained between 71 and 83 percent of the total variance in the data (the second

---

* For a variety of reasons, development and procurement of two of the three systems studied have been halted; neither of these two systems (the baseline or prototype versions of the LOS-F-H and Aquila RPV) is expected to be fielded.

factor revealed never accounted for more than 1% of the variance). The results of this initial analysis supported the view that the four workload scales essentially provide assessments of a single common factor.

Table 1

Magnitude and Source of the "OWL Factor"

STUDY	Magnitude	Source
LOS-F-H NDICE	79.6	TLX/OW/MCH/SWAT
LOS-F-H Generic SME	82.6	TLX/OW/MCH/SWAT
Crew	75.9	TLX/OW/MCH/SWAT
LOS-F-H FDTE Basic	79.4	TLX/OW/MCH/SWAT
LOS-F-H FDTE 48 Hour	81.5	TLX/OW
Aquila FDTE	75.2	TLX/OW/MCH/SWAT
Aquila FIREX	83.4	TLX/OW
UH60A Simulator	71.4	TLX/OW/MCH/SWAT

During the second stage of the factor validity analyses, Jackknife PCAs were conducted of the workload measures in order to evaluate the **factor validity** or the stability of the factor loadings of the four scales. (The factor loading of each scale is the correlation of the workload scale rating values with the corresponding OWL factor scores.) For example, in the LOS-F-H NDICE study, there were four factor loadings and 6 subjects which yielded a 4 (loadings) by 6 (subjects dropped) matrix. The data matrix resulting from this analysis was examined by conventional repeated measures ANOVA. The BMDP2V program (Dixon, 1983) was used to perform these ANOVA.

Table 2 shows the results of the factor validity comparisons for all four rating scale techniques in each study for which the comparisons can be made. The table presents, for each study, the ordered mean factor loadings. The horizontal line underscores factor validity value differences which were shown by subsequent pair-wise comparisons to be non-significant. From this table, it may be seen that TLX had the highest factor validity, i.e., the greatest correlation with the OWL factor score, for each of five studies over three different systems. Comparing the other three scales across all the studies, OW is next best, followed by SWAT and MCH.

Table 2

Factor Validity Scores Across Studies

STUDY	TECHNIQUE (Mean Factor Loading)			
LOS-F-H NDICE	TLX(.935)	OW(.927)	MCH(.862)	SWAT(.860)
LOS-F-H Generic	TLX(.924)	OW(.905)	MCH(.904)	SWAT(.778)
LOS-F-H Basic	TLX(.924)	SWAT(.900)	OW(.898)	MCH(.818)
Aquila FDTE	TLX(.910)	SWAT(.893)	OW(.869)	MCH(.833)
UH60A Simulator	TLX(.899)	OW(.872)	SWAT(.805)	MCH(.799)

Operator Acceptance

Another source of comparative information on the four rating scales was the reactions of the operators to the scales. The usability or acceptance of any operator reporting instrument is a critical (if not the critical) selection criterion. This dimension is of interest because the increased operator acceptance of a "subjective" measurement tool may result in increased willingness to express a valid opinion that can be taken seriously and used.

After using the four rating scales to rate all the mission segments and operator tasks of interest in three separate studies (i.e., the LOS-F-H NDICE study, the RPV FDTE study, and the UH-60A simulator study), a rating scale questionnaire was administered which solicited judgments regarding the procedures and tests instruments, particularly those used to measure OWL. The questionnaire asked the subjects to compare the four OWL rating scales and indicate the following:

- Which one they liked best,
- Which one was the easiest to complete,
- Which one was the hardest to complete, and
- Which one allowed the best description (rating) of the workload that had been experienced.

The administration of this questionnaire facilitated an open discussion of the four workload assessment scales.

Table 3 shows the number of times each scale was given the highest ranking for each of three different systems separately for each acceptance criterion. It may be seen that the majority of subjects both liked TLX the best and believed that it provided the best description of workload. Subsequent follow-up interviews revealed that many who thought TLX, with its six component dimensions, provided the best description, liked it best for that reason.

Regarding the relative ease of use, most subjects thought OW the least difficult to complete and almost all indicated the MCH was the hardest to complete. Follow-up interviews revealed that ease of completing the OW scale lead some subjects to judge it as allowing the best description of workload. Not solicited from the subjects, but freely offered by most, were complaints regarding the difficulty of the special card sort procedure which is required before using SWAT (see the next two sub-sections).

Resource Requirements

Along with factor validity and operator acceptance, it is also important for practical purposes to know the relative resource requirements for utilizing the workload

Table 3

Operator Acceptance of Workload Rating Scales

Study	Rating Scale TLX	OW	MCH	SWAT
Which of the rating scales did you like the best?				
LOS-F-H	2	2	1	1
Aquila	7	3	3	1
UH-60A	5	7	2	2
Which rating scale was the easiest to complete?				
LOS-F-H	1	4	1	0
Aquila	3	4	0	0
UH-60A	2	11	2	1
Which rating scale was the hardest to complete?				
LOS-F-H	0	1	3	2
Aquila	2	0	8	2
UH-60A	3	2	9	5
Which rating scale do you think best allowed you to describe (or rate) the workload you experienced?				
LOS-F-H	5	0	1	0
Aquila	10	5	2	0
UH-60A	8	4	1	4

Note. Data shown are the number of times each scale is given the highest ranking.

assessment scales (i.e., how much does it cost to use each scale). Since each of the four rating scales is most likely to be used as a paper-and-pencil technique, there is little variation among them in material needs. The differences among the scales are reflected in time requirements (i.e., the time required for scale preparation, training or instructing raters to use them, completing the scales when they are administered, and analyzing the results obtained with the scales).

The time to complete or fill out each of the four types of scales was measured during the LOS-F-H NDICE study. The results of that effort are shown in Table 4. It is clear that TLX, with its six subscales, takes the most time to complete, while OW takes the least; the SWAT and MCH scales have intermediate mean completion time values.

The other time requirements were not systematically measured, but our experience is that the OW scale requires substantially less time to prepare, train or instruct, and analyze results than the other three scales. Much less data are generated in the unidimensional OW scale than the multidimensional TLX and SWAT scales, and the procedure for completing the OW scale is much simpler than the highly structured and relatively complex MCH scale. The multidimensional TLX and SWAT scales require

Table 4

Time (seconds) to Complete Workload Rating Scales

Scale	n	Mean	SD
TLX	38	51.3	29.5
OW	33	9.8	8.4
MCH	27	29.1	26.3
SWAT	27	33.6	24.6

more time for data reduction than the unidimensional OW and MCH scales. Finally, TLX and SWAT scales require additional analysis time to develop composite scores; SWAT requires a computer and TLX only a calculator to perform this task.

Special Procedures

The requirements for using the two multidimensional scales -- TLX and SWAT -- include some special procedures. These procedures are designed to elicit judgements from the raters concerning their perceptions of the relative salience of the scale component dimensions, independent of the workload ratings themselves. Of course, these special procedures require additional time. The SWAT technique requires a card sorting procedure in which the rater determines the rank order of all possible combinations of the three levels of each of its three dimensions of workload. The TLX technique requires a paired comparison procedure for its six dimensions to determine individual weightings of each dimension's importance to workload, separately for each rated task.

We obtained data on the time required to complete these special procedures in only one study -- from the six soldiers used in the LOS-F-H NDICE study. The times to complete the SWAT sort procedure were 25, 30, 33, 34, 43, and 45 minutes (mean = 35 minutes). The times required to complete the TLX paired-comparison procedure were approximately 6-7 minutes for the first task and 2-3 minutes for subsequent tasks. The additional information gained from the multidimensional representation of workload may bear the cost of the additional time required for these special procedures. However, in the case of TLX, our research suggests its special paired comparison procedure may be skipped without compromising the measure (see a later sub-section of this part of the report, or Byers, Bittner, & Hill, 1989).

While most subjects were able to perform the TLX paired comparisons procedure correctly and with no apparent difficulty, the same cannot be said for the SWAT card sorting procedure. The required SWAT procedure not only takes a substantial amount of time to complete, but also presents other problems for some of the subjects. More specifically, 23 (or 43%) of 54 subjects performing the SWAT card sort did not initially produce adequate SWAT card sorts. The unsuccessful subjects produced inconsistent

sorts with excessive axiom violations according tc the SWAT User's Manual (Armstrong Aerospace Medical Laboratory (AMRL), 1987). Our observations suggest that the problem may be more pronounced for less verbal and less "sophisticated" subjects. Consequently, time must be set aside for resolving such problems, though we have encountered subjects for whom such problems could not be resolved. In the latter case, the experimenter must also be prepared to either use the data from these subjects despite their inconsistent SWAT card sorts or discard them.

Summary of the Direct Comparison Among Empi ical Scales

The results presented in the preceding four sub-sections tempt one to conclude that TLX was the most acceptable and usable workload assessment scale, and that MCH was the least acceptable scale. This conclusion must, of course, be moderated by the knowledge that there was a limited subject sample size and a limited span of test conditions in the present set of workload assessment studies.

If time is a major consideration, the data presented in Table 4 show that TLX individual assessments required more time-to-complete than the other measures. However, if factor validity or operator acceptance are the major criteria, Tables 2 and 3 show that TLX is superior to the other measures. Except for the more than 5-fold time-to-complete of TLX relative to OW, these time-to-complete differences may be judged relatively marginal in the context of other time costs (e.g., the time to perform an analysis of video tape recordings). However, given the moderately high factor validity of OW across all of these studies, arguments may be made for its use (vs. TLX) for screening very large numbers of mission segments with respect to overall workload (e.g., in preparation for more diagnostic evaluation of "workload problem areas"). These arguments, it is noteworthy, are predicated on tradeoffs of temporal cost, scale validity, and subject availability factors which may be evaluated only on a case-by-case basis.

Based on the results of all of our investigations and our review of the literature, the present authors have concluded that TLX is generally the preferred workload rating scale for all but screening applications, in which case it may be appropriate to use OW.

General Efficacy of Empirical Workload Rating Scale Techniques

The general usefulness, efficacy, and validity of workload rating techniques were further examined in terms of their ability to capture changes in the workload imposed upon an operator by the system, mission, and environment. The dependent measure used for these analyses was as often as not the **OWL factor score**. More specifically, the issue is not which one of the workload rating techniques provided useful information and it is not if a specific technique yielded useful information. Rather, the larger issue is if operator ratings of workload provide useful information. The **OWL factor score** is used to evaluate this issue since, when two or more different rating scale techniques are used in a given study, it is the best possible estimate of whatever is being measured, in common, by those techniques. In some cases, because of their demonstrated factor

validity and operator acceptance, only one or two scales -- the TLX or OW -- were used to assess workload. In these cases, either **factor** scores derived from the ratings of just these two scaling techniques or the ratings from just one technique were used in the workload analyses.

In succeeding subsections, the results of seven different types of workload analyses are summarily presented and discussed. These seven types of analyses address the following issues:

- The relationship between workload ratings and system performance,
- The sensitivity of workload ratings to expected variations in imposed or experienced workload,
- The effect of extended-duration missions on workload ratings,
- The use of subject matter experts to augment the workload ratings of small populations of experienced operators,
- The effects of delays in workload ratings,
- The information value of TLX subscale ratings, and
- The necessity to weight TLX subscale rating to derive a global workload rating value.

Workload Ratings and System Performance

It could be argued that if workload rating data are to have any practical value, they must impact on the decision processes which drive Army programs. To do this, the proponent of the program needs to be convinced that those data relate to the desired outcome of that program. This would certainly seem to be true in the case of a materiel systems development program. Consequently, any effort to validate workload rating scale techniques must demonstrate that the data they produce are related to the performance of the system. This dimension of validity is often called **criterion-referenced validity**, where the criterion of success for a system is its capability to correctly perform mission essential functions.

Workload studies on the LOS-F-H and UH-60 systems yielded different results about the relationship between operator workload and system performance. (No measures of system performance were available for the two Aquila RPV studies.) For the LOS-F-H system, step-wise regression analyses were conducted to examine this relationship. In the NDICE study the dependent variable was the system performance scores of 0, 1, or 2 (based on the number of targets destroyed during an engagement opportunity) and the independent variable was the TLX ratings of the system operators. In the FDTE Basic study the dependent variable was a performance score determined by

the percentage of successful engagements over all passes and missions and the independent variable was the **OWL factor scores** of only the ROs. In both studies, dichotomous independent variables were also used to index the operator making the rating. The results of these analyses are summarized in the regression lines given in Figures 1 and 2. As may be seen, increases in operator workload were associated with decreases in system performance. In both studies, the multiple correlations were significant, R = 0.66, and 0.65, respectively.

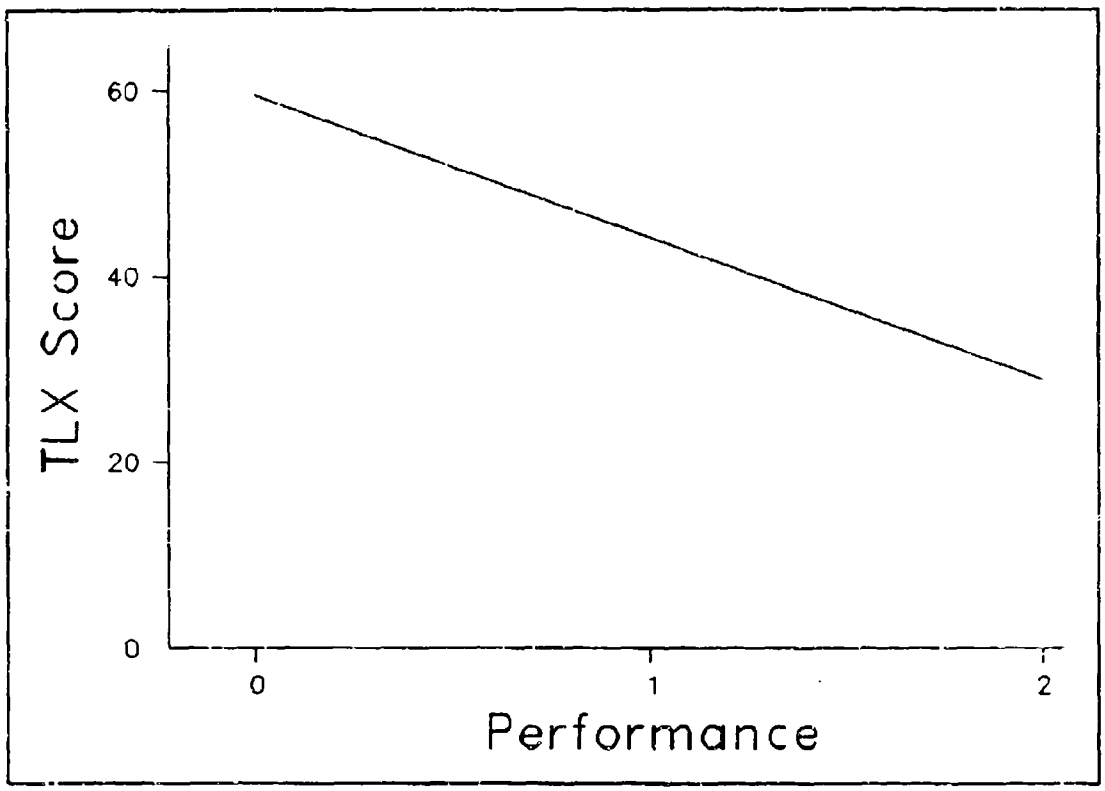


Figure 1. *The relationship between T'LX workload ratings and system performance in the LOS-F-H NDICE study.*

For the UH-60A study, an independent observer present during the simulator flight rated the performance of the pilot and copilot for the required tasks in each mission segment (e.g., performing the necessary navigation subtasks while enroute from a pickup zone to a landing zone). The pilot and copilot provided near real-time ratings of overall workload (OW) and peak workload (PW) for these same tasks. Analyses of these data revealed no significant relationship between the ratings of crew performance and the workload ratings of the crew members. Contrary to the two LOS-F-H studies, the UH-60A performance measures were based not on the performance of the system, but on the performance of the operators. One would think that the workload experienced by the operators would be more closely linked to the operator-based performance data than the system performance; the latter is also driven by factors unrelated to the operator's performance.

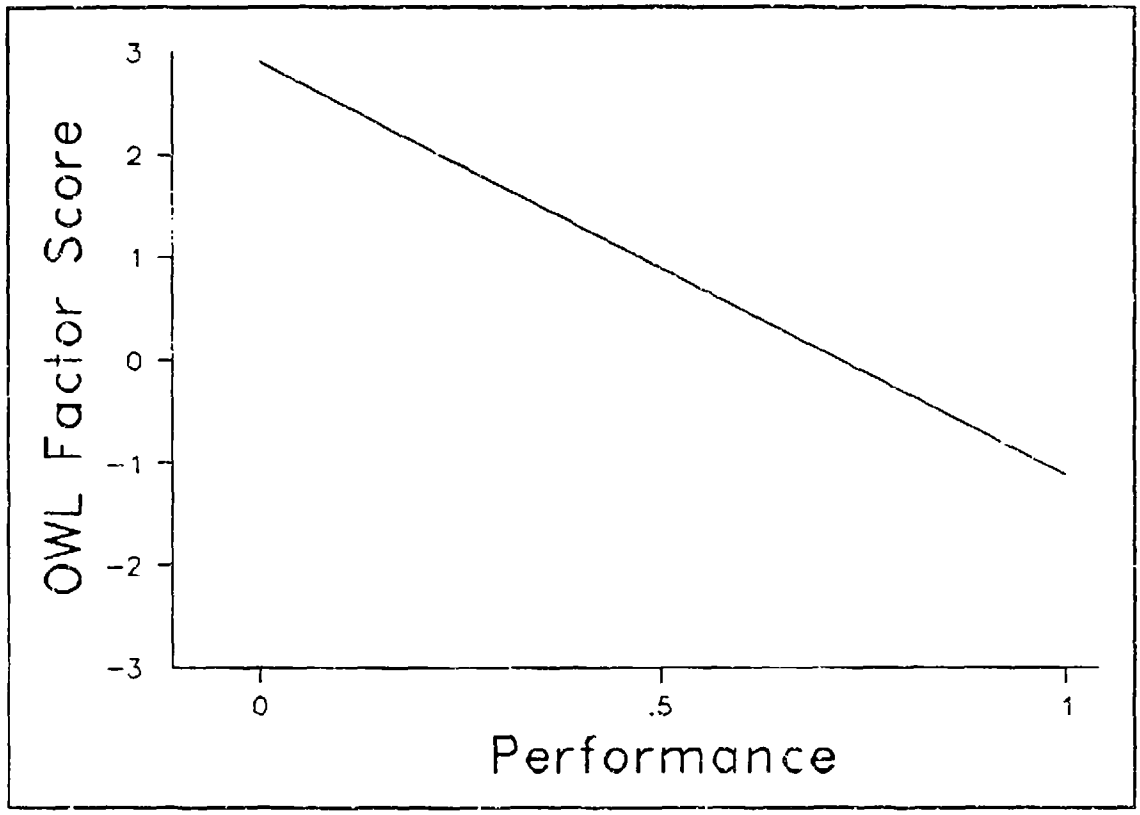


Figure 2. *The relationship between OWL factor scores and system performance in the LOS-F-H FDTE Basic study.*

The absence of a relationship between ratings of UH-60A crew performance and the workload ratings of the crew members may be attributed to the method employed to rate performance. The scale used to rate performance is the same as the one which is used to evaluate the performance of student pilots. The subjects in the UH-60 study, however, were all very experienced instructor pilots. It is quite likely that these subjects could perform at uniformly high levels regardless of workload levels. It is also quite likely that the performance rating scale designed for use with undergraduate pilots was simply not sensitive to the high levels of performance expected of instructors.

In summary, it is possible to demonstrate a meaningful **quantitative** relationship between workload ratings and system performance, even up to several months following the events to be rated. However, the presence of this relationship will depend upon the procedures used to measure both variables.

Sensitivity to Expected Variations in Imposed OWL

The analysis of factor validity described in an earlier section showed that the **OWL factor scores** are sensitive to variations in the aggregated data from each study. In other words, the **OWL factor scores** were able to discriminate among and quantify different levels of task loadings. It is noteworthy that the sensitivity of the workload

rating measure is meaningful in a practical sense as well. Admittedly, the workload ratings obtained in these empirical studies generally did not reveal any surprises. Their most important contribution was their capability to quantify the expected but not well differentiated differences in the amounts of workload that would be imposed upon and experienced by the operators of systems. These quantified values of workload may be shown to vary as a function of mission conditions, crew duty assignments, and characteristics of the test situations.

The sensitivity of the ratings to imposed workload was established for all three systems studied and in all but one of the OWL studies.* Succeeding paragraphs illustrate the types of measurement sensitivity that were found for each of the three systems.

LOS-F-H. The FDTE Basic study results revealed a significant interaction between mission segment and crew member position, as illustrated in Figure 3. The driver (DR) reports less than average workload in all segments. The radar operator (RO) and electo-optics operator (EO) report higher than average workload for the acquisition/tracking and reload mission segments. However, during the emplacement segment of the mission, the RO has higher than average workload average while the EO has much lower than average workload. More detailed analyses showed that the acquisition/tracking workload was primarily attributable to high mental demand while that for the reload segment was due largely to physical effort. Hence, the workload ratings are clearly sensitive to various workload components, including both cognitive and physical aspects.

The LOS-F-H studies also revealed a significant interaction for workload ratings as a function of operator tasks and types of targets. As illustrated in Figure 4, the Generic study showed that dual targets were associated with higher workload than single targets, and target identification (ID/IFF) and track-to-intercept tasks have higher workload than target handoff tasks. These results are in line with operational expectations. However, OWL differences also were seen in the interaction between target type and operator tasks. For the handoff and tracking tasks, dual-target passes were rated higher in workload than were single-target passes. For the identification tasks, on the other hand, dual rotary wing engagements had higher workload ratings than either dual fixed-wing or single rotary-wing engagements. Thus, for the identification task, both the number and type of target seem to affect workload. Dual rotary-wing aircraft may pose greater workload for the identification task due to unpredictable nature of the typical flight path (e.g., close-in, pop-up).

---

* The exception, the LOS-F-H NDICE study, found no stable relationships for workload ratings as a function of specific mission segments or target conditions. In that study, large variations in workload ratings were observed across subjects; and these clouded statistical comparisons of the mission segments and test conditions of interest. It appears that those ratings, made after watching video recordings of their own performance, reflected idiosyncratic differences in the mission conditions that were being rated. Even though the "same" types of missions and mission segments were selected for each operator to rate, variations in the actual conduct of these missions caused them to in fact be different from one another in terms of their impact on the operators. Hence, the failure to find stable relationships for workload ratings in this study may, in part, be due to the sensitivity of the ratings to variations in task loadings.

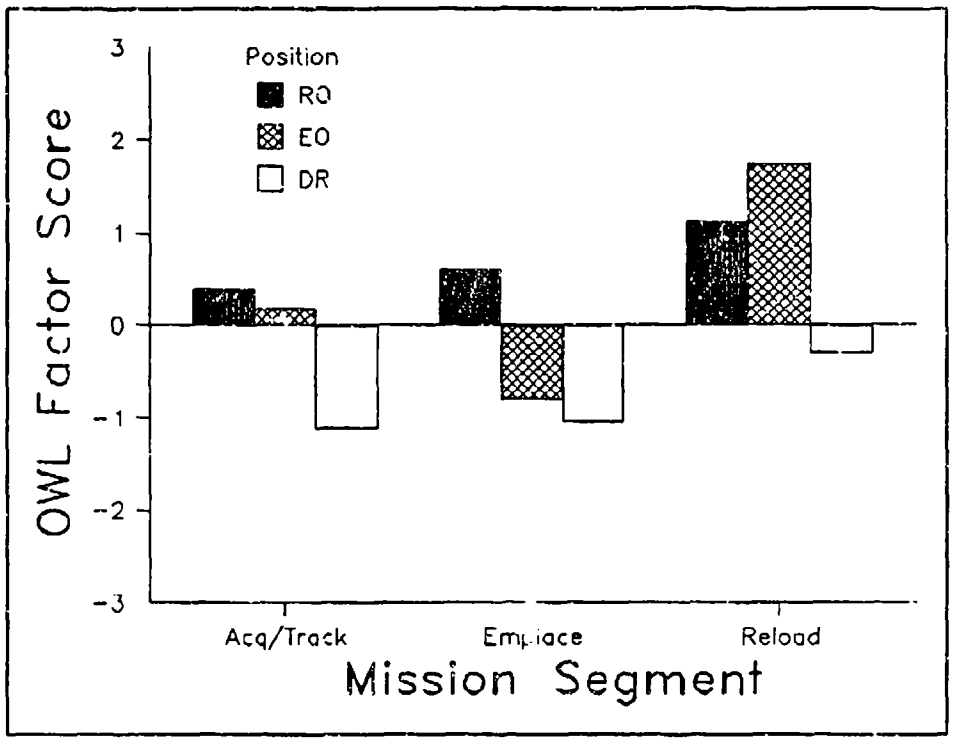


Figure 3. *The effect of mission segment and crew member position on workload in the LOS-F-H FDTE Basic study.*

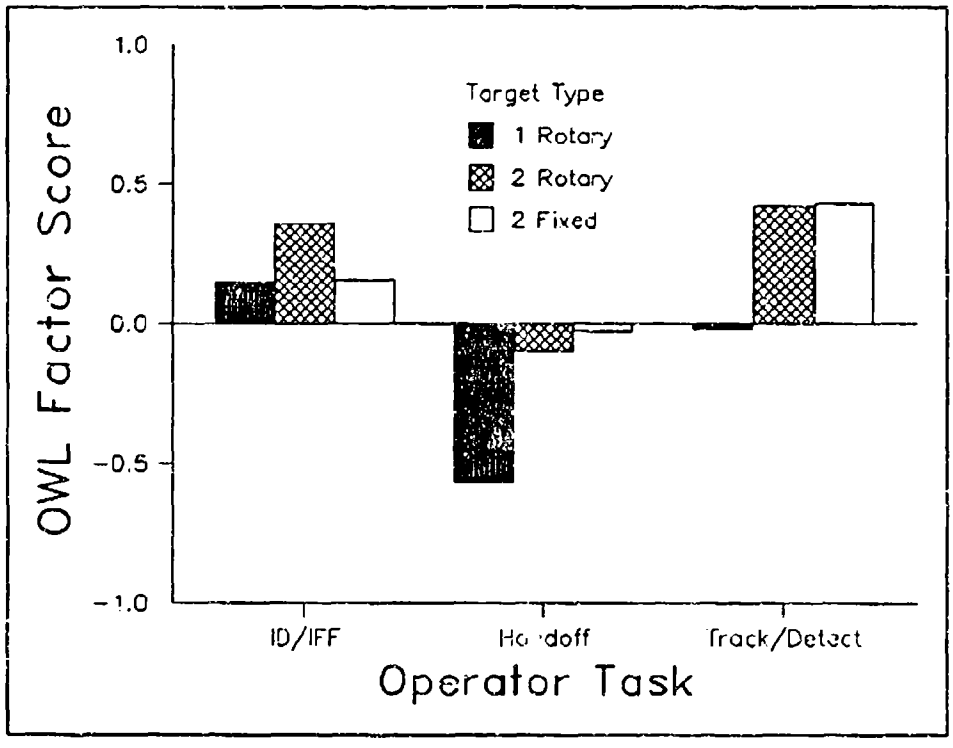


Figure 4. *The effect of operator task and target type on workload in the LOS-F-H Generic study.*

Aquila RPV. The FIREX 88 study results also revealed a significant interaction between mission segment and crew position, as illustrated in Figure 5. It may be seen that while the mission commander (MC) has the highest and relatively consistent OWL factor scores, the workload ratings of the AVO and MPO vary considerably and in opposite directions over segments. These results make sense. The workload of the MC is driven by a fairly constant level of responsibility over an entire flight of the RPV. The MPO has no direct responsibility during launch and recovery when the mission payload is not in use but higher than average workload during the flight when the mission payload is used to perform mission essential functions. On the other hand, the AVO has the least workload in the flight segment of an RPV mission when flight operations are relatively routine but higher than average workload during launch and especially during recovery when various problems can and often do arise.

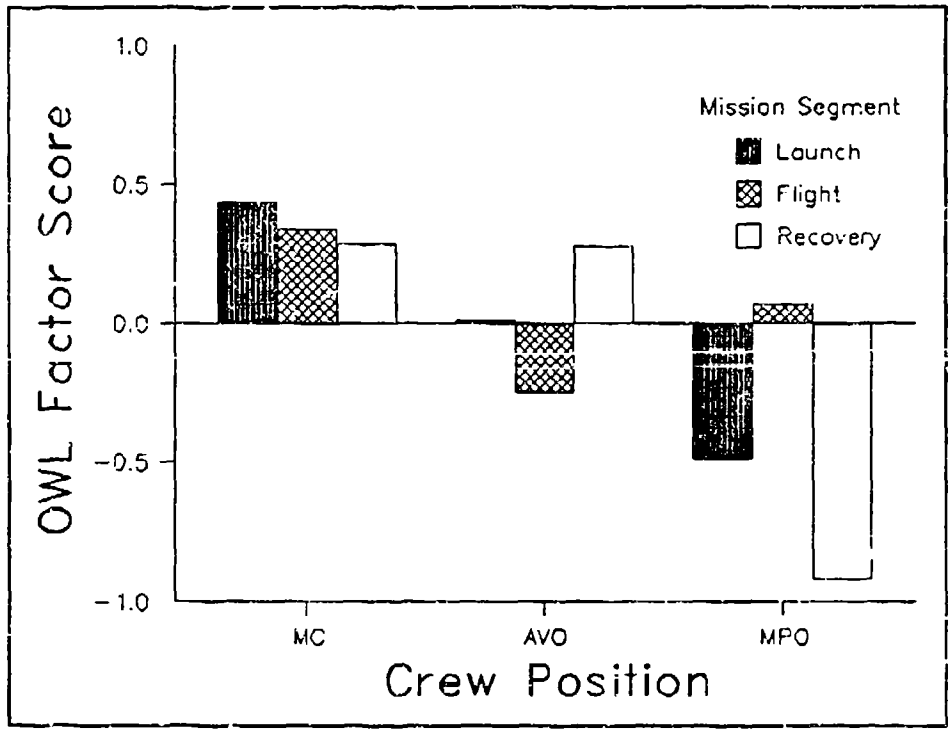


Figure 5. *The effect of mission segment and crew member position on workload in the Aquila FIREX 88 study.*

Figure 6 shows the workload experiences of Aquila crew members as a function of the contrast in test conditions between the OT II and the FDTE. The main effect of test conditions reflects the reduced workload in the FDTE due to several factors, including: new improved search software, intensified training, a more restricted scope of the mission, and the fact that the air vehicle was not actually flown but mounted to the underside of a manned aircraft.

The significant interaction between workload setting (FDTE and FIREX 88) and Aquila crew position is illustrated in Figure 7. It was to be expected that the AVO would have a higher level of workload in FIREX than in the FDTE (he actually flew the

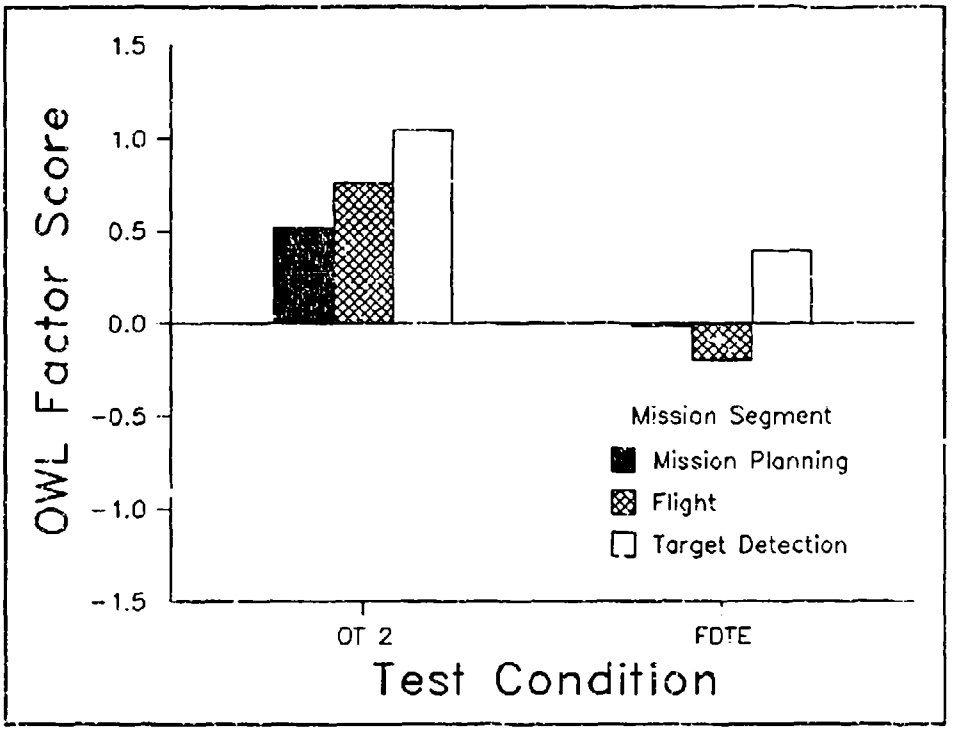


Figure 6. *The effect of test condition and mission segment on workload in the Aquila RPV ground control station.*

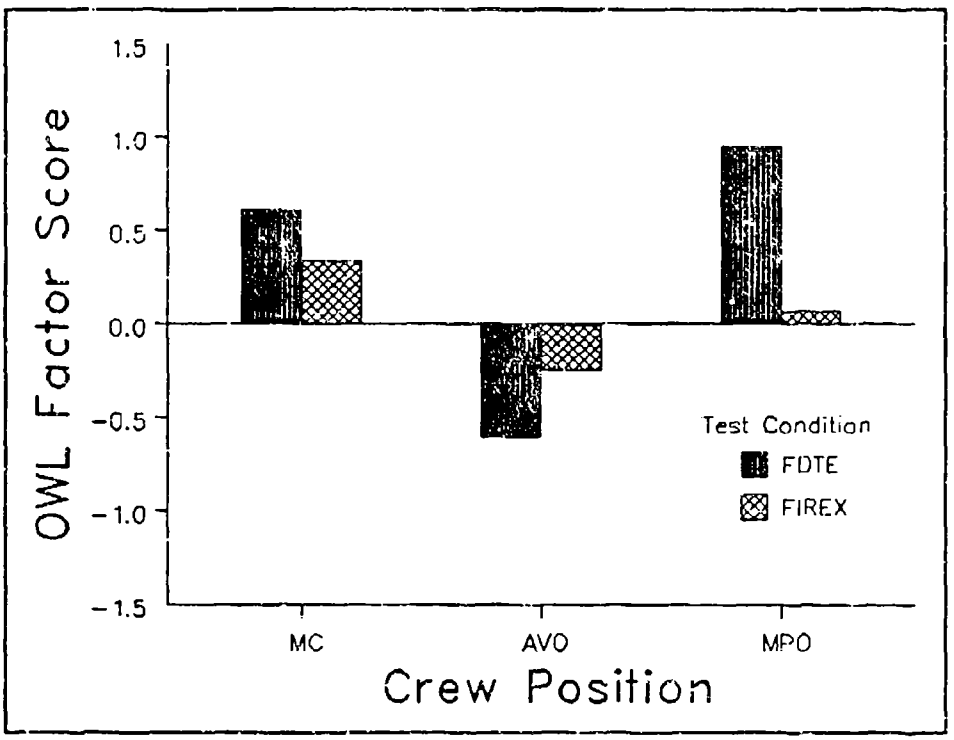


Figure 7. *The effect of test condition and crew member position on workload in the Aquila RPV ground control station.*

AV only in the FIREX). The opposite effects were expected for the MPO (because target detection was not a major objective of the FIREX flights) and the MC (because the MC in FIREX were more experienced than those in the FDTE and the pressure to maximize performance was reduced).

UH-60A. Workload ratings in the UH-60A study were also shown to be sensitive to variations in task demands. For example, the effect of different mission segments on mean real-time ratings of pilots and co-pilots is shown in Table 5. The greatest level of workload was found in Segment 12 in which an engine failure occurred enroute from the forward arming and refueling point (FARP) to the start point (SP). The least workload occurred during refueling operations at the FARP (Segment 11), and during the two initial flight segments enroute to the pickup zone.

Table 5

Mean Real-time Workload Ratings for Mission Segments in the UH-60A Simulation Study

Segment Number	Description	Code	Rating
1	Startpoint to Checkpoint 1	SP-CP1	3[illegible].0
2	Checkpoint 1 to Pickup Zone	CP1-PZ	38.4
3	Pickup Zone Operations	PZ Ops	42.5
4	Pickup Zone to Landing Zone	PZ-LZ	50.4
5	Landing Zone Operations	LZ Ops	46.3
6	Landing Zone to Pickup Zone	LZ-PZ	**
7	Pickup Zone Operations	PZ Ops	40.9
8	Pickup Zone to Alternate LZ	PZ-Alt LZ	49.5
9	Alternate LZ Operations	Alt LZ Ops	48.6
10	LZ to Forward Arming & Refueling Point (FARP)	LZ-FARP	**
11	FARP Operations	FARP Ops	31.5
12	FARP to Special Including Engine Failure	FARP-SP	52.9

Note. Segments 6 and 10 are not included due to missing data.

Analysis of Extended-Duration Missions

One of the goals of the OWL Program was to investigate how workload changes over an extended period of time. This issue is important because real-world missions are often extended over long durations. Furthermore, workload effects which are not apparent during short, discrete tasks may be cumulative and produce overload conditions only after an extended period of time. Figure 8 shows the mean workload rating of each of two crews as a function of time into their respective 48-hour missions. It may be seen that workload ratings generally increase across time for both crews.

Since task demands were relatively constant over the duration of the 48-hour mission, the increase in workload over time may reflect a decrease in the resources the system operators have to commit to mission essential tasks. In this case, workload may be associated with fatigue, which would be expected to increase over time during continuous operations. This does not necessarily mean that the ratings represent merely

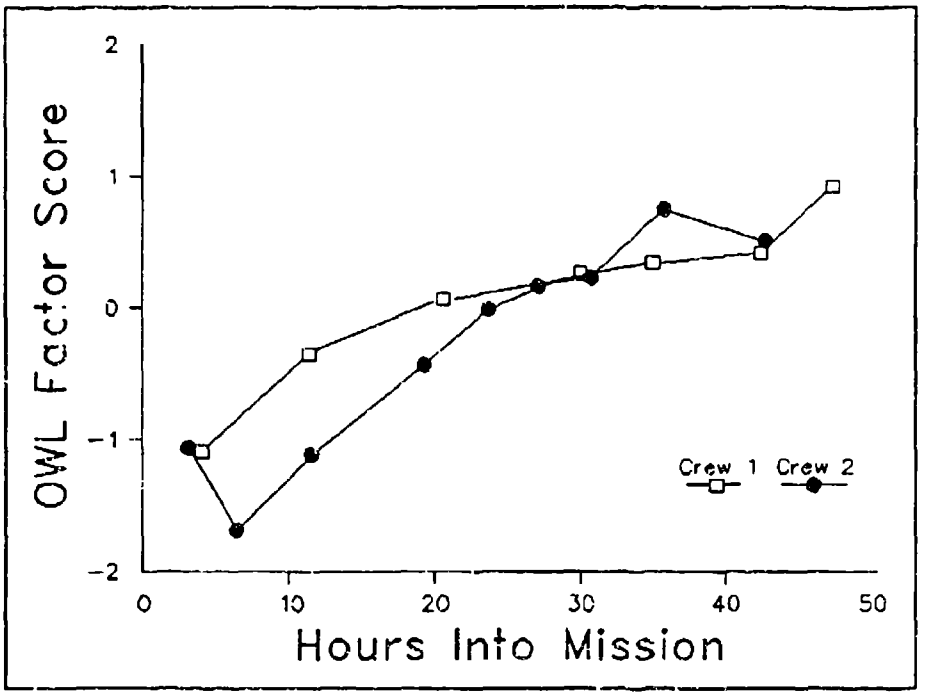


Figure 8. *The effect of an extended duration mission on workload in the LOS-F-H FDTE 48-hour mission study.*

a cumulative score which would have to increase over time. An alternative possibility is that operators "averaged" workload for the mission up to the point where the ratings were obtained. Though the general trend in the data was increasing, there where several points at which the mean ratings decreased, lending support to the second interpretation.

Analysis of Rater Experience with the System

In one of the OWL Program primary studies, i.e., the LOS-F-H Generic study, we had the opportunity to compare the workload ratings of experienced crew members with those of other SMEs for descriptions of generic mission segments. This comparison is important for two reasons. First, there are often very few well-trained soldier/operators (especially on a new or prototype system) and their availability is usually restricted. Small samples limit the utility of statistical analysis techniques and prevents wide generalization of the results. SME participation in workload analyses would be one way to augment the population of subjects. Second, there is the question of what constitutes a well-trained subject. A comparison of the workload ratings of highly trained crews and clearly less well trained SMEs would permit an analysis of the impact of rater experience level. Table 6 shows the diverse backgrounds of the SMEs used in this study.

The results showed that crew members of the LOS-F-H system and SMEs generated essentially equivalent OWL factor scores across the conditions of the generic missions. Most importantly, the two groups showed the same orderings of workload ratings over conditions for the two measures with the highest factor validities (i.e., TLX

Table 6

Experience of SMEs in the LOS-F-H Generic Study

SME	ASSOCIATION WITH SYSTEM	INVOLVEMENT IN NDICE	TRAINING ON SYSTEM	WATCHED FILMS OF NDICE (10 OR MORE)	OTHER AIR DEFENSE EXPERIENCE	MILITARY EXPERIENCE
1	MANPRINT	YES	YES	YES	YES	YES
2	MANPRINT	NO	YES	YES	NO	YES
3	MANPRINT	YES	YES	YES	YES	YES
4	MANPRINT	YES	YES	YES	YES	YES
5	TRAINING	YES	YES	YES	YES	YES
6	TRAINING	NO	NO	YES	YES	YES
7	TRAINING	NO	YES	NO	YES	YES
8	TRAINING	NO	YES	NO	YES	NO
9	TRAINING	NO	NO	NO	YES	YES

and OW). These results suggest that SMEs may be successfully used to augment a limited operator pool of subjects in making workload ratings of **generic missions.** It would be a mistake, however, to assume that any SME could make workload ratings equivalent to those of an experienced operator. Clearly, caution is advised until alternative criteria for defining SMEs are defined and evaluated.

Effects of Delays in Rating

One of the stated advantages of the operator rating techniques is their non-intrusiveness. By deferring the workload measurement response until after a task has been completed, these techniques permit the task to be performed with minimum interference. The other side of the coin is that when the operator does make his or her response, **memory** is called into play; the operator must remember the situation and the experiences associated with it in order to make a workload rating. This, in turn, raises questions about how to interpret the rating. If the ratings are not made in real time, does memory distort the judgment of workload? What are the effects of different task-response time lags on workload ratings and what is the source of these effects? The OWL Program did not conduct a controlled study of the effect of delays in ratings, so a definitive answer to these questions cannot be obtained from the collected data. However, during the course of the OWL Program, different time lags were used in different studies. Consequently, several salient observations can be made about this issue.

Subjects were quite able to provide workload ratings with substantial face validity even when the time lag was large, as in the LOS-F-H NDICE and Generic studies (see Figures 3 and 4). In the NDICE study, the time lag was about 10 weeks and subjects were asked to rate very specific mission segments. In that study, however, the use of video recordings helped the subjects to recall their experiences from those segments, thus easing the memory burden. For the Generic study, the same subjects and the same types of mission segments were rated, and the time lag was about six months. In this study, another procedure eased the dependence on memory -- the use of verbal descriptions of "generic" or average mission segments under the given set of task conditions.

During the UH-60A study using the flight simulator, different and much smaller time lags were used. In particular, the subjects provided ratings in near "real time" -- at the first acceptable time following the mission segment of interest -- and "post time" -- following the completion of the entire mission. The corresponding time lags were much shorter (i.e., seconds or minutes compared to weeks or months). As was the case in the two LOS-F-H studies cited in the preceding paragraph, subjects were able to provide reasonable OWL ratings for the mission segments using both real- and post-time ratings of workload. It should be noted, however, that the values of the real-time ratings were greater than those of the post-time ratings (46.0 and 41.0, respectively). We speculate that during the mission real-time ratings of workload were elevated relative to post-mission ratings due to the uncertainty and anticipation of mission tasks remaining to be completed.

Finally, it could be argued that the LOS-F-H Prospective study represents instances of "negative" time lag in which the ratings were made to a task planned to occur at some future point in time. As will be described later in this results section, the results of that study showed that operators could make reasonable ratings to the extent that their general knowledge encompassed a situation similar to that being rated. When this was true, the rating situation was similar to that of the generic study, in that the subjects are mentally picturing themselves in a given situation or mission based upon their general knowledge, and making their ratings using that mental picture.

Analysis of TLX Subscale Ratings

An important distinction among the four workload rating techniques selected for analysis in the OWL Program is the **information output** of the scales. The OW and MCH scales produce a single overall judgment of workload for each rated situation, while the TLX and SWAT techniques produce component subscales information as well as overall judgments. This subsection addresses the nature, analysis, and interpretation of subscale information provided by multidimensional workload assessment techniques. In particular, it deals with the issue of the **diagnosticity** of these techniques.

Diagnosticity refers to the extent to which the specific source or cause of workload is revealed by the measurement technique. Workload techniques may be diagnostic in that they may be used to identify the potential components (e.g., mental, physical, and temporal) which contribute to the perception of workload. The essence is to be able to identify the specific mechanism or process involved during the performance of a particular task under particular conditions, especially if that process is overloaded.

Because of resource limitations, the OWL Program focused on limited subscale analysis of just one of the two multidimensional scales. The TLX was selected because of its consistently higher factor validity and operator acceptance. As described in Appendix A, the TLX subscales are: mental demand, physical demand, temporal demand, performance, effort, and frustration. Overall workload is calculated as a weighted average of these six subscale ratings. The TLX subscales were analyzed for the

LOS-F-H (Generic, Basic, and Prospective), Aquila FIREX, and UH-60A studies. Table 7 shows the grand mean ratings of each subscale for the operators of three different systems: Aquila GCS, UH-60A, and (prospectively) LOS-F-H. The main effect of subscale ratings was significant for each of these three systems. It may be seen that there were similarities and differences in subscale values across the systems. In each, mental demand is the greatest and physical demand nearly always the smallest contributor to workload ratings. In terms of differences, for example, frustration is not a major contributor to workload for the UH-60A and LOS-F-H prospective ratings but is the second greatest contributor to workload in the Aquila ratings.

Table 7

Mean Weighted TLX Subscale Ratings for Three Different Systems

	TLX Subscale					
System	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration
LOS-H-F Prospective (all cases)	142.7	11.7	98.0	94.2	94.8	56.7
Aquila FIREX (GCS only)	189.8	14.3	140.9	98.5	84.0	129.2
UH-60 Blackhawk (all cases)	114.8	40.1	112.3	61.9	108.6	31.5

Analysis of interactions of workload subscale values with key independent variables of a study provides even more useful information than an analysis of only main effects. Changes in the pattern of subscale values across mission segments, duty position, and target configuration can help to identify workload problems and their sources at a finer level of detail than can a main effect. An example of a two-way interaction between mission segment and subscale values was found for the UH-60A study and is illustrated in Figure 9. It may be seen that three factors contribute to the higher mean workload shown for the Pickup Zone to Landing Zone (PZ to LZ) mission segment compared to the other three; the PZ to LZ segment has larger effort, physical, and mental components than the other three segments. This result is reasonable since the PZ to LZ segment consisted of flying through hostile territory carrying a heavy load, a situation in which the platform can become quite unstable.

A three-way interaction involving TLX subscales is illustrated in Figure 10. This figure shows the effect on total weighted workload scores of different sources of workload in various mission segments (Acquisition/Tracking [Acq/Track], Emplace, and Reload) and crew members in different duty positions (radar operator [RO], electro-optics operator [EO], and Driver [DR]). For example, during acquisition/tracking, the RO experiences more total workload than the EO (though not significantly more), although the EO experiences more temporal demand than the RO. Another example is that the RO always has higher performance subscale ratings (i.e., he perceives he has been less successful in accomplishing his task) than either the EO or Driver.

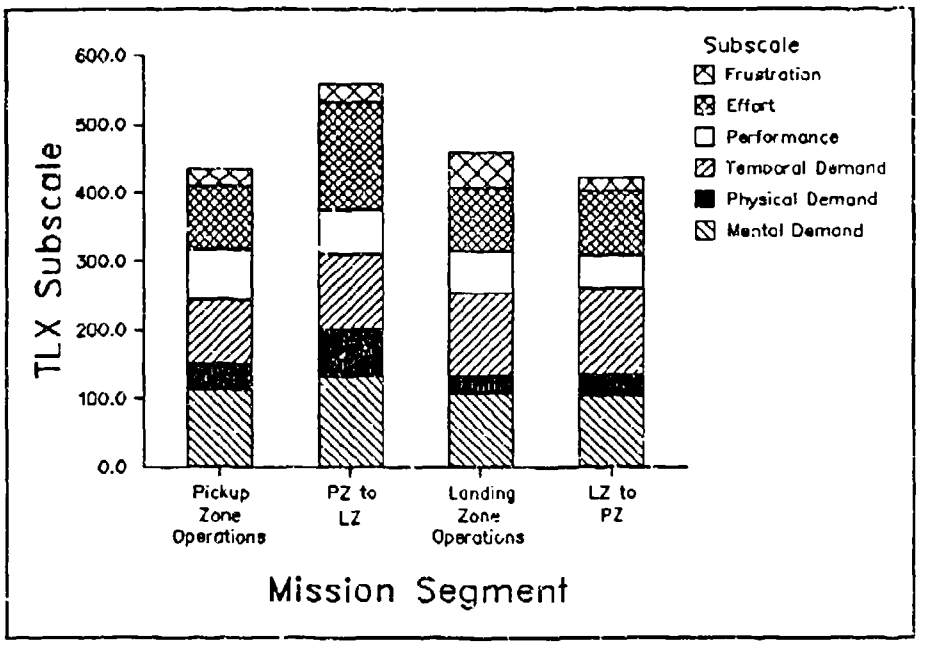


Figure 9.
The effect of mission segment and TLX subscale on weighted subscale scores in the UH-60A simulator study.

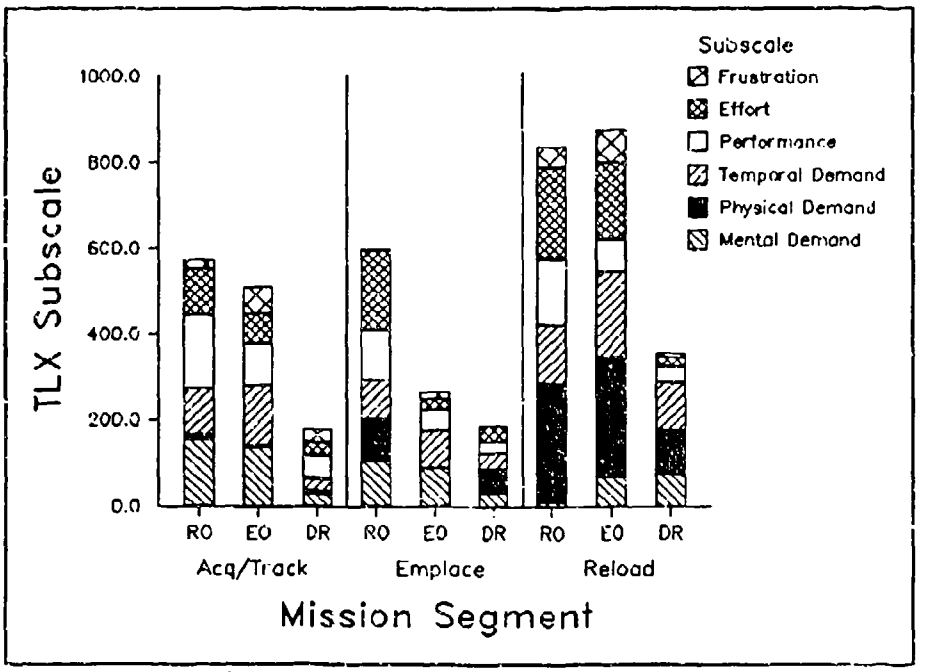


Figure 10. *The effect of mission segment, crew member position, and TLX subscale on weighted subscale scores in the LOS-F-H Basic study.*

Comparison of Raw and Weighted TLX Ratings

As described above, multidimensional rating techniques such as TLX and SWAT require separate procedures designed to address individual differences in the perception of factors (i.e., sub-scales) which contribute to overall workload experiences. However, these special procedures are both cumbersome and time consuming. Because of the potential utility of these multidimensional techniques, it would be desirable to reduce the burden associated with or to eliminate entirely the need to use the special procedures. There has been some interest in using SWAT without the need for the card sort procedure (Biers & McInerney, 1988). However, since the OWL Program has shown clear advantages of the TLX over SWAT in both factor validity and operator acceptance, we focused on the TLX rating technique.

The standard TLX composite or global workload score is computed by multiplying each subscale rating by a weighting factor derived from the pair-comparison of all subscales for each task to be evaluated. These weighted ratings are then averaged to obtain the global score following procedures given by Hart and Staveland (1987). These authors stated that the weighted composite score produces more stable overall workload scores (i.e., scores with a smaller variance) than a rating obtained using the OW technique (which yields directly a single overall judgment of workload). This is a reasonable finding from a strictly statistical point of view. Because the standard composite TLX score is presumed to be the sum of (approximately) independent and identically distributed variables, it would have a smaller variance than a unitary score.

However, these authors did not compare TLX with an appropriate **unweighted** average of subscale rating values. If there were no paired comparison of TLX subscales, there would be no derived weights, and a "Raw TLX" (RTLX) could be calculated by simply averaging the subscale values, thus skipping the weighting step in both the experimental procedure and in the analysis. We calculated RTLX and compared it to TLX (and OW as a baseline) across a number of the OWL studies (see Byers, Bittner, & Hill, 1989). Table 8 summarizes the results of these comparisons. It may be seen that RTLX has slightly lower mean values and slightly **lower variability** than the TLX, and a **very high correlation** with TLX (averaging 0.977 across five studies). The assertion that TLX scores would have less variability than OW scores was confirmed. Hart (personal communication, October, 1989) offered an explanation of our results. She noted that our findings are reasonable for complex, realistic tasks whose workload is due to the contributions of several subscales. However, she also argued that for simple, "unitary" tasks whose workload is principally due to a single subscale (i.e., the types of tasks more typical of the laboratory rather than the field), the equality of RTLX and TLX may not hold. The rationale behind this interpretation of our results should be further explored.

Evaluation and Validation of Analytical Techniques

A major premise that continually recurred throughout the duration of the OWL Program is that analytical or predictive workload assessment techniques can be extremely

important in influencing the development of a system. Appropriate use of these techniques allows the human factors analyst or practitioner to make meaningful contributions early in the design phase of an emerging system. Such early involvement not only would improve the quality of the emerging product design, but would also lay the groundwork for continuing useful workload contributions throughout later phases in the system development process.

Table 8

Comparison of OW, TLX, and Raw TLX (RTLX) Workload Scores Across Studies

Study	n	OW	TLX	RTLX	r
LOS-F-H NDICE	72	37.71 (26.03)	36.78 (22.35)	34.00 (21.14)	0.982
LOS-F-H GENERIC	230	57.26 (26.51)	52.25 (21.53)	50.21 (22.18)	0.967
LOS-F-H Basic	204	35.20 (20.47)	31.23 (17.55)	28.96 (16.42)	0.981
PMS FDTE	66	46.36 (22.58)	36.15 (18.99)	36.50 (17.84)	0.960
AQUILA FIREX	105	46.48 (27.96)	43.00 (24.15)	38.75 (21.95)	0.973
ACROSS ALL STUDIES	677	45.80 (26.27)	41.27 (22.43)	38.97 (21.81)	0.977

Note. The values shown for the ratings are the mean and standard deviation. The PMS FDTE refers to a workload study conducted on another system, the Pedistal Mounted Stinger (see Byers, 1989).

As valuable as the analytical workload assessment techniques are, they suffer from two disadvantages in most applications to date: coarseness and lack of validation. The coarseness of the outputs of analytical methods is not really a disadvantage -- it is more a property of the early stages of the system development process. At early stages, little firm system information is typically available and is usually of a very general nature. No assessment technique can produce finer-grained output than that of the input information.

Validation of analytical workload assessment techniques is complex and difficult, involving both technical and resource problems. We were fortunate in the OWL Program to have the opportunity to apply two different analytical techniques, one each to two different systems. Prospective TLX ratings were used to assess the opinions of experts toward the workload that would be associated with some proposed changes to the LOS-F-H system. Workload "predictions" made with the TAWL/TOSS technique were developed and matched to empirical, real-time workload ratings for the UH-60A helicopter. Each of these two applications of analytical techniques is described below.

Prospective Application of TLX Ratings

The LOS-F-H Prospective study examined the workload ratings of operators for **hypothetical** situations, including more realistic target configurations, new radar equipment, more realistic employment of multiple fire units, and different task allocation among crew members. There were two types of validity sought in this study.

First and more modest was the desire to establish **face validity**, or to answer the question, "do the quantified prospective ratings reflect reasonable relationships between workload and the variables of interest?" To address this validation objective, the mean prospective rating results were discussed with system and tactics experts who, on the whole, judged them to be reasonable.

A clear example of the face validity of prospective ratings are the results obtained for prospective ratings of more realistic target configurations. The mean TLX ratings for the "average" number of aircraft that had been experienced during a one-hour acquisition/tracking segment in the LOS-F-H FDTE Basic study and "double" that number were 38.7 and 46.2, respectively. Likewise, the mean ratings for the "typical" rotary-wing or fixed-wing attack and a hypothetical attack simultaneously by two fixed-wing **and** two rotary-wing aircraft were 31.7 and 45.8, respectively. These results confirm the expectation that the serial nature of the RO and EO tasks in an engagement sequence may lead to easy handling of single targets, but potential problems when multiple targets appear in rapid succession. An equally relevant example of this first

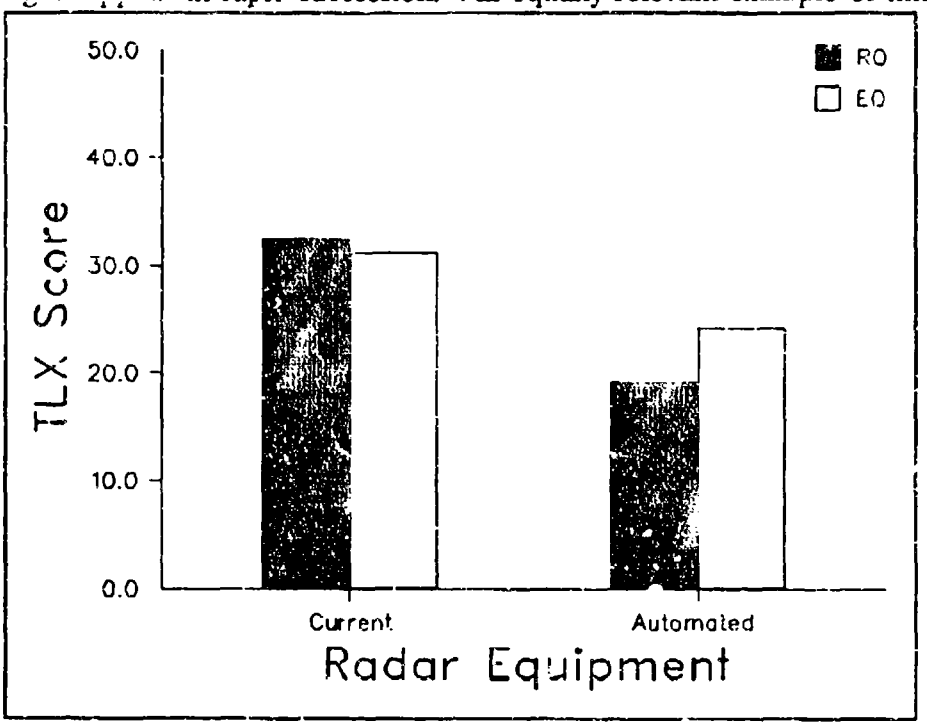


Figure 11. *The effect of proposed automated radar equipment and crew member position on TLX ratings in the LOS-F-H Prospective study.*

type of validity is illustrated in Figure 11. This figure shows that a proposed new radar which would automate tasks such as those associated with target identification, classification, and engagement priority would reduce the workload for both the RO and EO, but more so for the RO than the EO.

A second and more ambitious validation objective was to establish **predictive validity**. It was our plan to participate in the next LOS-F-H test opportunity, the FDTE - Phase II, to empirically evaluate some prospective ratings made at the conclusion of the LOS-F-H FDTE studies we participated in for this report. Unfortunately, a shift in the FDTE - Phase II schedule made it impossible to fulfill these plans. We had been particularly interested in testing workload predictions about the effects of multiple fire units and a reallocation of crew responsibilities, both of which were to occur for the first time during the Phase II field test. Figure 12 illustrates the prospective ratings associated with a more realistic configuration of several fire units. The "master fire unit" is the one with an active radar, which receives command and control data over an active radio network, and which determines the assignment of targets to fire units. The slave vehicle is responsible for engaging the assigned targets. Figure 3-12 illustrates a significant interaction of operation mode (Master, Slave, and Autonomous) and duty position (RO and EO). The overall workload of the RO and EO is rated about the same in the Autonomous Mode. However, the RO is projected to experience greater levels of workload than the EO in the Master Mode and the reverse is projected to occur in the Slave Mode.

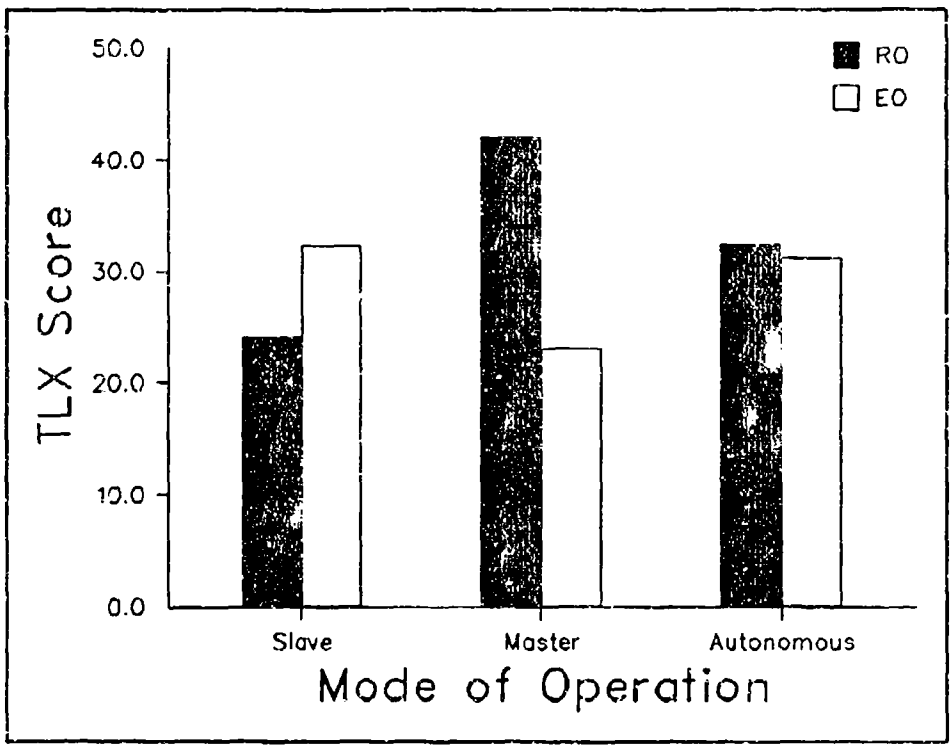


Figure 12. *The effect of proposed mode of operating multiple fire units and crew member position on TLX ratings in the LOS-F-H Prospective study.*

The prospective operator rating method is most likely to be effective for hypothetical situations where the operators have had some relevant, similar personal experiences. Such a situation would allow the operators to have a "mental anchor" for their prospective judgments. One set of prospective ratings used for this study -- that for the new organization of tasks across crew members -- did not seem to have a comparison base. The proposed reorganization would place the senior crew member, who served as a mission commander and squad leader, in the driver's position in the fire unit. From that location, this crew member would keep the fire unit in the air battle and monitor the ground battle. He would maintain direct contact with the air defense platoon leader and with the maneuver force that the fire unit was assigned to protect, and would drive the fire unit from one battle position to another. The individuals assigned to the RO and EO positions would serve the duties normally assigned to these two positions. Essentially, in the proposed organization, the mission commander no longer functions as the RO but instead as the driver.

Figure 13 illustrates an interaction effect on prospective ratings by crew position for the current and proposed organizations. As has been described previously, Figure 13 shows that in the current organization, workload of both the RO and EO exceeds that of the driver, especially for more difficult missions. In contrast, in the proposed organization, while the workload projected for the driver **plus** mission commander/squad leader position is higher than for the driver in the current organization, that for the RO and EO is only marginally affected and, in fact, tends to decrease for more difficult missions. In summary, all three positions are predicted to have essentially the same level of workload in the proposed organization. However, soldiers indicated that the proposed organization appeared very strange, largely because "drivers" are generally the lowest ranking soldier in most, if not all, Army land vehicles. We speculate that the absence of familiarity (and perhaps some hostility) with the proposed organization reduced the size of the effects found.

The prospective application of the TLX operator rating technique also produces significant findings with respect to the TLX subscales. For example, Figure 14 illustrates a significant three-way interaction involving the mode of operation of multiple fire units, crew member duty position, and TLX subscale. It may be seen from this figure that "performance" subscale rating is larger for the RO in the Slave mode of operation than for any other duty position by mode combination, suggesting that an individual in the slave-RO position will perceive he has been relatively unsuccessful in performing his task. It may also be seen that the mental and temporal demands for the Master-RO is much greater than for the Master-EO, while the mental and temporal demands for the Slave-EO is greater than for the Slave-RO. This latter observation is in line with expectations.

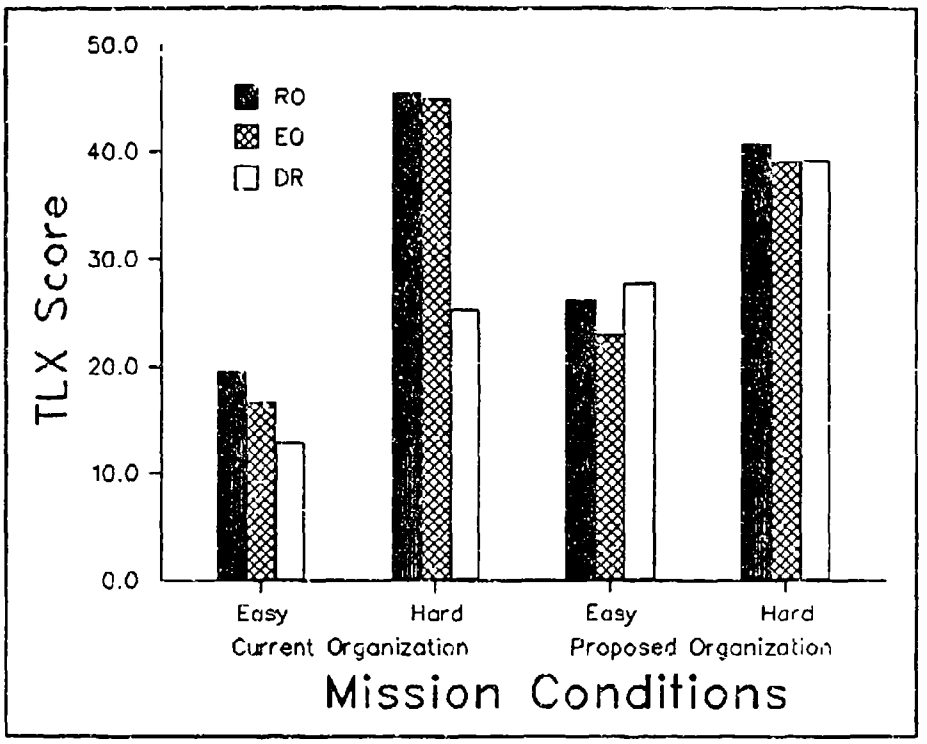


Figure 13. *The effect of crew organization, mission difficulty, and crew member position on TLX ratings in the LOS-F-H Prospective study.*

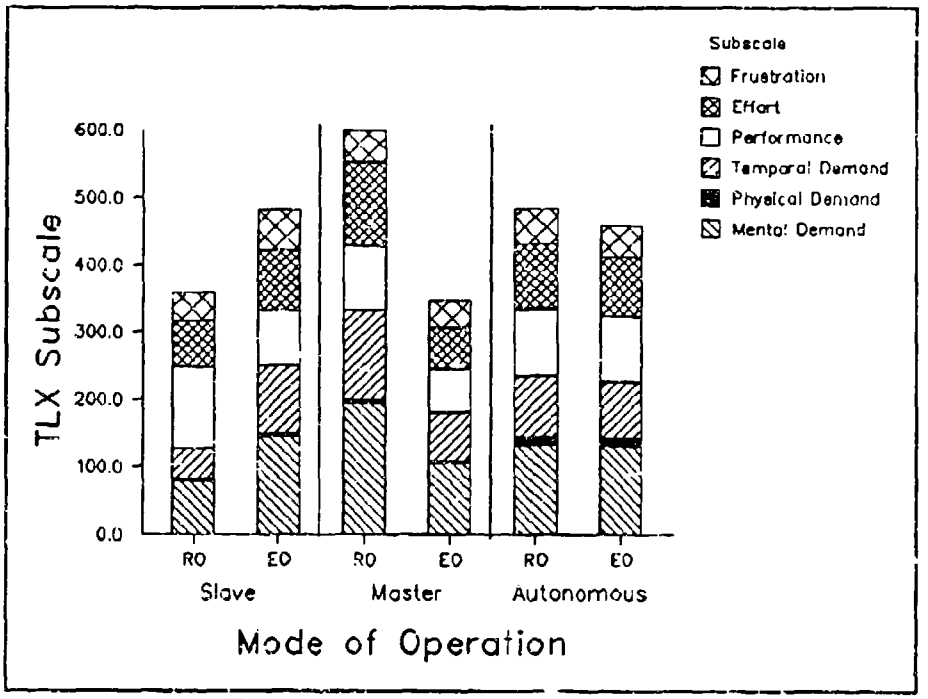


Figure 14. *The effect of mode of operating multiple fire units, crew member position, and TLX subscale on weighted subscale scores in the LOS-F-H Prospective study.*

Analysis of the TAWL/TOSS Methodology

The TAWL/TOSS methods were developed over several years by Anacapa Science and ARI (Bierbaum et al., 1990). This portion of the OWL Program is of particular importance because it is the only study within the program which examined the validity of a task analytic and simulation method.

It is not straightforward to define what constitutes validation of an analytical model (such as TAWL/TOSS) that predicts complex human behavior, especially if the model output is a construct (such as "workload") which has many alternative definitions. For the UH-60A study, the ability of TAWL/TOSS to reasonably track changes in the workload (as rated "real-time" by pilots and copilots throughout a mission) was analyzed. In adhering to the OWL Program objectives to provide useful assistance to Army developers, it is not important to determine if the predictions precisely match empirical data. Rather, it is important to determine if TAWL/TOSS can provide reliable (if approximate) indications of potential workload problems.

Required TAWL/TOSS inputs include a detailed task analysis with low-level task times, channel-specific workload ratings for each of the low-level activities, and a set of scenario decision rules that drive the simulated operator's task selection. Using this information, TAWL/TOSS generates a timeline of low-level activities at fixed half-second intervals. To determine the channel workload at each half-second interval, the TAWL\TOSS model sums the workload estimates across tasks that are concurrently performed at that time. If the sum of any component channel (e.g., visual) exceeds 7 within a half-second interval, an overload is defined to have occurred for that channel in that interval.

The purpose of the current study was not to investigate prediction of "overload" by the TAWL\TOSS model. Rather, the study focused on validating the underlying workload database and the scenario generation rules developed specifically for the TAWL\TOSS UH-60A model. The approach used was to compare real-time operator ratings of workload with TAWL/TOSS-based predictions of workload. For example, techniques were devised to derive predictions of real-time Overall Workload (OW) ratings from the output provided by the TAWL/TOSS model. This technique proved to be quite reasonable. A significant correlation was found across crew members between TAWL\TOSS-derived predictions of OW and the real-time OW ratings ($r = 0.82$). This high correlation suggest the validity of the underlying TAWL/TOSS data base and scenario generation techniques.

Figure 15 illustrates this finding graphically by mission segment, separately for the pilot and copilot. As may be seen in the figure, TAWL/TOSS predictions track the relative overall workload between segments. However, the real-time OW ratings tend to be higher than the TAWL/TOSS-based OW predictions. The one exception to this trend - Pilot data for the first mission segment shown in the figure - suggested the

possibility of a special modeling problem (D. B. Hamilton & C. R. Bierbaum, personal communication, January, 1990). If the pilots' data for this segment are removed, the correlation rises to $r = 0.95$, with the relationship accounting for 90 percent of the total variance in the data.

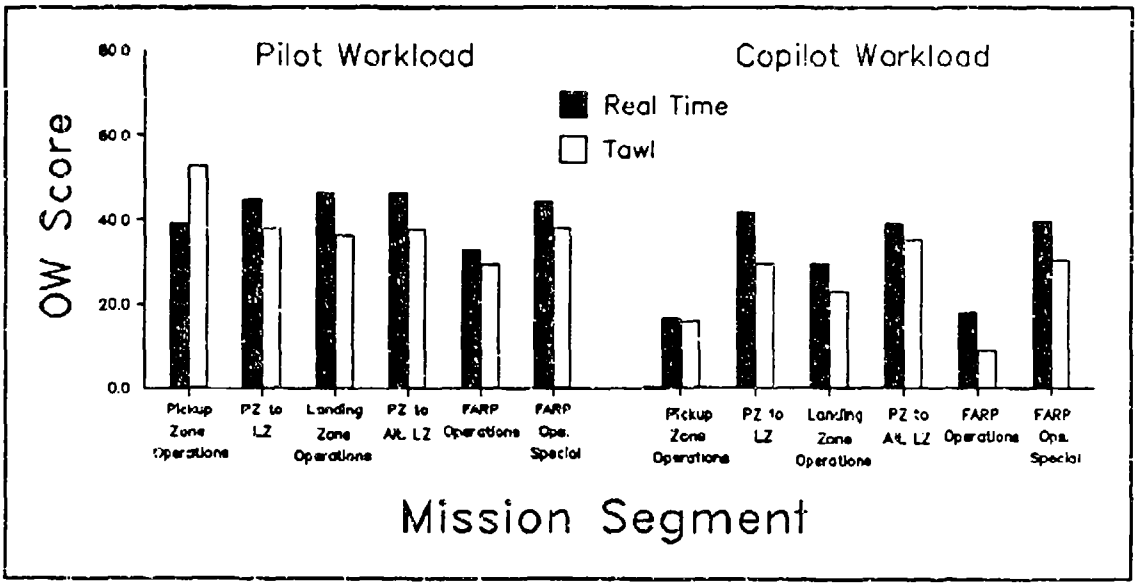


Figure 15. *The effect of mission segment and crew member position on real-time ratings and TAWL/TOSS model predictions of overall workload in the UH-60A simulator study.*

SUMMARY AND CONCLUSIONS OF THE OWL PROGRAM PRIMARY RESEARCH STUDIES

This report describes the methods and procedures used, and the findings obtained from a series of eight separate studies across three Army systems. It addresses the application of both empirical methods for evaluating the workload associated with the operation of Army systems and analytical methods for predicting that workload. It presents and discusses the results obtained from these studies in terms of their meaningfulness or validity for a number of different practical topic areas.

The empirical methods examined were four operator rating techniques: TLX, SWAT, OW, and MCH. In the studies reported, TLX was consistently highest in factor validity and operator acceptance. For these reasons, TLX is recommended for all but screening applications, where OW (because of its simplicity and convenience) may be used as a first step. The empirical workload ratings are shown to be sensitive to changes in system performance and in the expected levels of workload imposed upon the operator by the system, mission, and operational conditions. Additional analyses show that the ratings are robust with respect to delays between a workload experience and its rating, and to variations in rater experience with the system under consideration. The TLX subscale ratings are shown to contain potentially useful information concerning the source or cause of experienced workload. Finally, if experimental resources are limited, the raw average of TLX subscale ratings are shown to produce composite or global workload scores essentially equivalent to those obtained using the standard weighted average of TLX subscale ratings.

The analytical methods studied were prospective operator ratings using the TLX scale and the TAWL/TOSS task analytic and simulation model. The prospective rating technique shows promise as a method for identifying potential workload problems in emerging systems. The TAWL/TOSS model is shown to have a capability to track empirical workload ratings. This indicates that the TAWL/TOSS model also has potential as an analytical workload estimation technique that may be used to predict workload early in the system development process. More research is indicated to fully exploit these analytical techniques.

Future Research Directions

Based on accomplishments and lessons learned from the recently completed OWL Program and from other related research programs, several areas for future work can be described. These include continuing work to generally improve our understanding of the concept of workload and its relationship to operator and system performance. In addition, research must proceed to identify cost-effective methods for reducing the impact of excessive OWL on soldier, system, and unit-level performance effectiveness.

In terms of our understanding of workload and its relation to performance, these areas of research should be pursued:

- Validation studies on an expanded set of workload assessment methodologies as they apply to a larger class of systems operating in more diverse environments. The database that addresses workload assessment techniques is too limited in scope.

- Further improvements of our capabilities to assess operator workload issues during system front-end analysis. Clearly, improved analytical techniques are required to predict workload early in system development where the greatest design flexibility is available with the least impact on system cost.

- Development of a more complete understanding of the effects of workload on human performance by expanding our research to include instances of "underload" as well as overload, and, perhaps more important, the performance consequences of transitions between these two extreme levels of workload.

- Better understanding of how workload analyses can be used to diagnose the sources of workload extremes. In spite of the current availability of "multidimensional" assessment techniques, it is not at all clear that we can adequately diagnose the cause of a workload problem for a system designer.

- Improvement of the ability to assess, understand, and utilize differences among individual soldiers in their reactions to workload extremes. It is generally understood that individual differences exist, but there is little research to relate them to workload.

- Development of a means to quickly incorporate new knowledge generated by the types of research described above into an expert system such as OWLKNEST. The capability to specify the relative values of various operator assessment techniques is important at all stages in the process of system development. The advice supplied by this expert system should be validated by application to real systems.

In terms of developing cost effective solutions or countermeasures to workload extremes, two different but obviously interrelated types of research are needed:

- Methods need to be developed for actually decreasing the extremes in workload imposed upon soldiers. These methods may be based on the design and development of (a) the hardware/software system and its interface with the operator; (b) the organizational unit within which the system is placed; or (c) the operational tactics, techniques, and procedures used during employment of the system.

- Methods are needed for increasing the soldiers' capability to successfully cope with extremes in operator workload. These methods may draw upon: (a) the identification, selection, and classification of soldiers whose performance is relatively tolerant to workload extremes, or (b) the design and implementation of training programs to develop effective individual and unit-level workload management strategies.

The need clearly exists for extending and enriching the total database that relates operator workload to soldier, system, and unit-level performance effectiveness. What remains to be determined is the ability to effectively and efficiently respond to that need. In part, the availability of required research support will determine the limits of our response. Our willingness and ability to change some basic orientations to developing research programs may be equally important.

REFERENCES

Armstrong Aeromedical Research Laboratory (AAMRL). (1987). Subjective workload assessment technique (SWAT): A user's guide. Dayton, OH: AAMRL, Wright-Patterson Air Force Base.

Bierbaum, C. R., Fulford, L. A., & Hamilton, D. B. (1990). Task analysis/workload (TAWL) user's guide - Version 3.0 (ARI Research Product 90-15). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A221 865)

Biers, D. W., & McInerney, P. (1988). An alternative to measuring subjective workload: Use of SWAT without the card sort. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1136-1139). Santa Monica, CA: Human Factors Society.

Bittner, A. C., Jr., Byers, J. C., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1989). Generic workload ratings of a mobile air defense system (LOS-F-H). Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1476-1480). Santa Monica, CA: Human Factors Society.

Bittner, A. C., Jr., Zaklad, A. L., Dick, A. O., Wherry, R. J., Jr., Herman, E. D., Bulger, J. P., Linton, P. M., Lysaght, R. J., & Dennison, T. W. (1987). Operator workload (OWL) assessment program for the Army: Validation and analysis plans for three systems (ATHS, Aquila, LOS-F-H). (TR 2075-3b). Willow Grove, PA: Analytics, Inc.

Byers, J. C. (1989). Workload assessment of the pedestal mounted stinger (PMS) (Technical Memo 7). Willow Grove, PA: Analytics, Inc.

Byers, J. C., Bittner, A. C., Jr., & Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? Advances in Industrial Ergonomics and Safety: Vol. 1. (pp. 481-485). London. Taylor and Francis.

Byers, J. C., Bittner, A. C., Jr., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1145-1149). Santa Monica, CA: Human Factors Society.

Byers, J. C., Christ, R. E., Hill, S. G., & Zaklad, A. L. (1988). Workload assessment of Aquila remotely piloted vehicle (RPV) operations during an operational exercise (Technical Memo 4). Willow Grove, PA: Analytics, Inc.

Byers, J. C., & Hill, S. G. (1989). Comparison of subjective workload ratings to field test performance of the LOS-F-H mobile air defense missile system (Technical Memo 8). Willow Grove, PA: Analytics, Inc.

Byers, J. C., Hill, S. G., Zaklod, A. L., & Christ, R. E. (1989). Aquila system report (Technical Report 2075-4a). Willow Grove, PA: Analytics, Inc.

Christ, R. E., Bulger, J. P., Hill, S. G., & Zaklad, A. L. (1990). Incorporating operator workload issues and concerns into the system acquisition process: A pamphlet for Army managers (ARI Research Product 90-30). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A228 489)

Cooper, G. E., & Harper, R. P. (1969). The use of pilot rating in the evaluation of aircraft handling qualities (NASA TN-D-5153). Moffett Field, CA: NASA-Ames Research Center.

Dixon, W. J. (Ed.). (1983). BMDP statistical software. Los Angeles, CA: University of California Press.

Eggleston, R. G., & Quinn, T. J. (1984). A preliminary evaluation of a projective workload assessment procedure. Proceedings of the Human Factors Society 28th Annual Meeting (pp. 695-699). Santa Monica, CA: Human Factors Society.

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), Handbook of perception and human performance: Vol. 2. Cognitive processes and performance. New York: John Wiley and Sons.

Harris, R. M., Hill, S. G., Lysaght, R. J., & Christ, R. E. (1992). Handbook for operating the OWLKNEST technology (HOOT) (ARI Research Note 92-49) and the accompanying software for the Operator workload knowledge-based expert system tool (OWLKNEST). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A253 412)

Hart, S. G., & Staveland, L. E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. S. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Hill, S. G., Byers, J. C., & Zaklad, A. L. (1989). LOS-F-H system report (Technical Report 2075-4b). Willow Grove, PA: Analytics, Inc.

Hill, S. G., Byers, J. C., Zaklad, A. L., Bittner, A. C., Jr., & Christ, R. E. (1988). Prospective workload ratings of LOS-F-H mobile air defense missile system (Technical Memo 2). Willow Grove, PA: Analytics, Inc.

Hill, S. G., Byers, J. C., Zaklad, A. L., & Christ, R. E. (1989a). Subjective workload assessment during 48 continuous hours of operations of the LOS-F-H. Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1129-1133). Santa Monica, CA: Human Factors Society.

Hill, S. G., Byers, J. C., Zaklad, A. L., & Christ, R. E. (1989b). Subjective workload ratings of the LOS-F-H mobile air defense missile system in a field test environment (Technical Memo 5). Willow Grove, PA: Analytics, Inc.

Hill, S. G., Lysaght, R. J., Bittner, A. C., Jr., Bulger, J. P., Plamondon, B. D., Linton, P. M., & Dick, A. O. (1987). Operator workload (OWL) assessment program for the Army: Results from requirements document review and user interview analysis (Technical Report 2075-2). Willow Grove, PA: Analytics, Inc.

Hill, S. G., Zaklad, A. L., Bittner, A. C., Jr., Byers, J. C., & Christ, R. E. (1988). Workload assessment of a mobile air defense missile system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1068-1072). Santa Monica, CA: Human Factors Society.

Hinkley, D. V. (1983). Jackknife methods. In S. Kotz, N.L. Johnson, & C.B. Read (Eds.), Encyclopedia of statistical sciences: Vol. 4 (pp. 280-287). New York: Wiley.

Iavecchia, H. P., Linton, P. M., Bittner, A. C., Jr., & Byers, J. C. (1989). Workload assessment during day and night missions in a UH-60 Blackhawk helicopter simulator. Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1481-1485). Santa Monica, CA: Human Factors Society.

Iavecchia, H. P., Linton, P.M., Harris, R. M., Zaklad, A. L., & Byers, J. C. (1989). UH-60 system report (Technical Report 2075-4c). Willow Grove, PA: Analytics, Inc.

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., Linton, P. M., Wierwille, W. W., Zaklad, A. L., Bittner, A. C., Jr., & Wherry, R. J., Jr. (1989). Operator workload: Comprehensive review and evaluation of workload methodologies (ARI Technical Report 851). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A212 879)

Masline, P. J., & Biers, D. W. (1987). An examination of projective versus post-task subjective workload ratings for three psychometric scaling techniques. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 77-80). Santa Monica, CA: Human Factors Society.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), Handbook of perception and human performance: Vol. 2. Cognitive processes and performance. New York: John Wiley and Sons.

Reid, G. B., Shingledecker, C. A., & Eggemeier, F. T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting (pp. 522-525). Santa Monica, CA: Human Factors Society.

Reid, G. B., Shingledecker, C. A., Hockenberger, R. L., & Quinn, T. J. (1984). A projective application of the subjective workload assessment technique. Proceedings of the IEEE National Aerospace and Electronics Conference (NAECON) (pp. 824-826). Dayton, OH.

Sheridan, T. B. (1980). Mental workload - What is it? Why bother with it? Human Factors Society Bulletin, 23, 1-2.

U. S. Army (1979). Human engineering requirements for military systems, equipment and facilities (MIL-M-46855B). Washington, D.C.: Department of the Army.

Vidulich, M. A., & Tsang, P. S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement application. Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.

Wierwilie, W. W., & Willeges, B. H. (1980). An annotated bibliography on operator mental workload assessment (Naval Air Test Center Report No. SY-27R-80). Patuxent River, MD: Naval Air Test Center, Systems Engineering Directorate.

APPENDIX A:

WORKLOAD ASSESSMENT INSTRUMENTS

CONTENTS

MODIFIED COOPER HARPER (MCH)

Description: The MCH is used to obtain ratings from 1-100 via a decision tree structure. Although derived from the Cooper-Harper, it was designed to be applicable to a broad number of operational environments (i.e., it is not specifically a pilot rating scale). It can be used in real-time operation.

Sensitivity: The scale has been reported to be sensitive to differences in task loading.

Diagnosticity: The MCH gives a global rating of workload.

Intrusiveness: Little, although it does require a judgment. There was concern (as with most subjective measures) that the judgment might interfere with flight duties, but ratings can be obtained real-time.

Implementation Requirements:

Data collection: Some method for collecting the ratings is needed -- either a 10 key pad or communications medium with which the operator can report the rating verbally. A copy of the scale for reference is also useful.

Operator training: The operators must be given an opportunity to become familiar with the rating scale, therefore some practice is necessary, although the scale is apparently easy to understand.

Operator Acceptance: The scale has been reported to be well received by experimental subjects who were pilots.

Safety: Plans must be made as to what to do if the operator is too busy to give a rating. Ratings should be secondary to the primary concern with operational safety (e.g., flying a plane or controlling a land vehicle).

Relative Cost of Use:

Testing time: Minimal.

Equipment: Minimal.

Setup and support: Minimal.

Data analysis: Descriptive and inferential statistics can be used. Graphical representations are useful. Caution is advised in assuming an interval scale, therefore non-parametric analysis may be more appropriate.

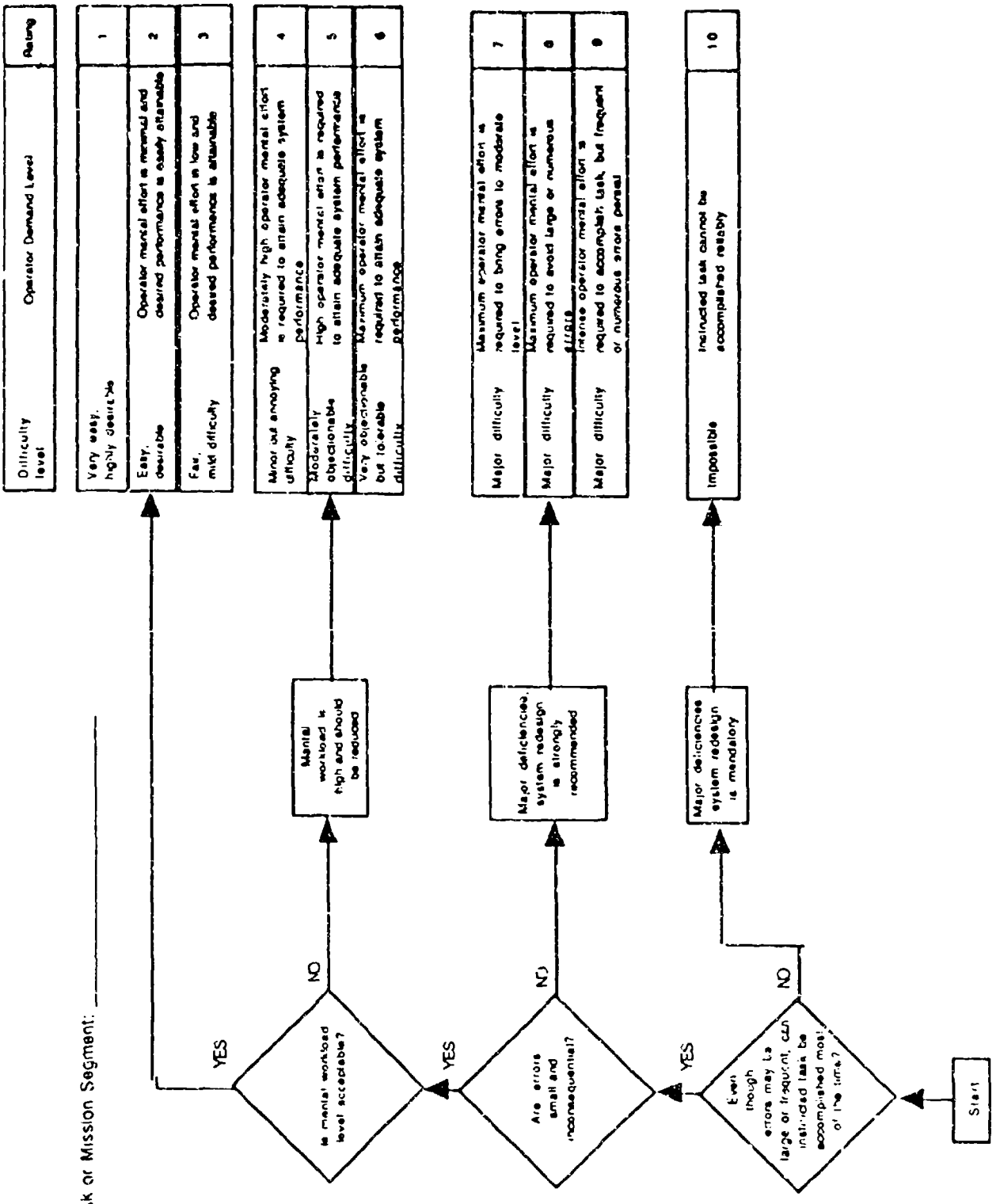
References:

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement application. Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.

Wierwille, W. W., Casali, J.G., Connor, S. A., & Rahimi, M. (1985). Evaluation of the sensitivity and intrusion of mental workload estimation techniques. In W. Roner (Ed.), Advances in man-machine systems research: Vol. 2 (pp. 51-127). Greenwich, CT: J.A.I. Press.

Wierwille, W. W., Skipper, J., & Reiger, C. (1984). Decision tree rating scales for workload estimation. Theme and variations (NASA-CP-2341). Proceedings of the 20th Annual Conference on Manual Control (pp. 73-84). Washington, D.C: NASA.

The MCH Scale



OVERALL WORKLOAD (OW)

Description: The overall workload (OW) scale is a unidimensional bipolar rating scale which an operator can use to give an absolute estimate of the workload experienced during a particular mission segment. The scale consists of a horizontal line divided into 20 equal intervals; the words "low" and "high" are placed, respectively, at the left and right ends of the scale. Numerical values, assigned by the analyst, range from 0 to 100.

Sensitivity: The scale has been shown to be sensitive to differences in task loading for a variety of different tasks, systems, and operational environments

Diagnosticity: OW gives only a global indication of the overall workload experienced by the operator.

Intrusiveness: Little, though it requires that the operator give an absolute judgment. Even so, studies have shown that OW ratings can be obtained in real time without interfering with the operator's performance.

Implementation Requirements:

Data collection: The OW scale can be administered during (real time), after (retrospectively), or before (prospectively) the operator performs the task of interest. The operator ratings can be obtained verbally, by paper and pencil, or electronically via a keypad.

Operator training: Some practice in using the scale and understanding the operational meaning of the scale (and of the concept of workload) is helpful.

Operator Acceptance: High

Safety: Plans must be made as to what to do if the operator is too busy to give a real-time rating. Normally, the analyst can ask for a retrospective rating at some period of time after the task of interest has been completed.

Relative Cost of Use:

Testing time: Minimal.
Equipment: Minimal.
Setup and support: Minimal.
Data analysis: Minimal.

Comments: When used retrospectively, after a long delay, the operator should be aided in recreating the experiences associated with the task when it was previously performed; audio and video recordings of task performance are helpful in this regard. When used prospectively, the operator or subject matter expert should be aided in creating a useful representation of the task as well as the system and operating environment which form the context of the task that is to be rated. In this latter case, the ratings of workload are made to descriptions of tasks and events that have not yet been personally experienced by the individual making the ratings (see Eggleston & Quinn, 1984).

References:

Byers, J.C., Bittner, A.C., Jr., Hill, S.G., Zaklad, A.L., & Christ, R.E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1145-1149). Santa Monica, CA: Human Factors Society.

Eggleston, R.G., & Quinn, T.J. (1984). A preliminary evaluation of a projective workload assessment procedure. Proceedings of the Human Factors Society 28nd Annual Meeting (pp. 695-699). Santa Monica, CA: Human Factors Society.

Hill, S.G., Zaklad, A.L., Bittner, A.C., Jr., Byers, J.C., & Christ, R.E. (1988). Workload assessment of a mobile air defense missile system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1068-1072). Santa Monica, CA: Human Factors Society.

Iavecchia, H.P., Linton, P.M., & Byers, J.C. (1989). Workload assessment during day and night missions in a UH-60 Blackhawk helicopter simulator. Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1481-1485). Santa Monica, CA: Human Factors Society.

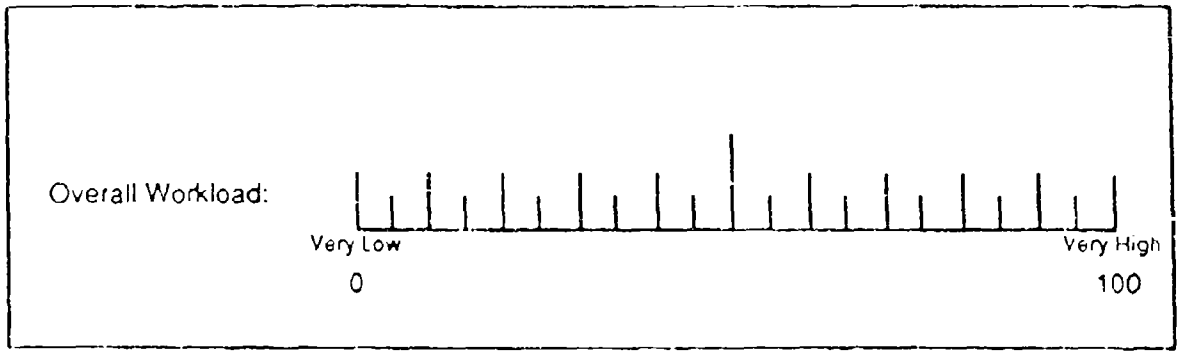
Vidulich, M.A., & Tsang, P.S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

AVAILABILITY: The OW scale is one of the subscales used during the construction of the TLX scale.

The OW Scale

Task or Mission Segment: _____

Please put a mark on the scale at the point which best corresponds to how you rate your overall workload.



SUBJECTIVE WORKLOAD ASSESSMENT TECHNIQUE (SWAT)

Description: SWAT uses the three dimensions of time load, mental effort load, and psychological stress load to assess workload. For each dimension, there are three operationally defined levels. SWAT has two parts: 1) a card sort procedure where the operator determines the rank order of all combinations of the three levels of the three dimensions; and 2) an event scoring part where the operator makes ratings of the three dimensions. Conjoint analysis is used to obtain a global workload rating between 0 and 100.

Sensitivity: SWAT has been demonstrated to be sensitive to task loading in a number of different types of tasks.

Diagnosticity: SWAT gives a global rating of workload. However, the three subscales can be examined individually and used for diagnostic purposes.

Intrusiveness: Little, although it does require a judgment. There was concern (as with most subjective measures) that the judgment might interfere with flight duties, but ratings were able to be obtained real-time.

Implementation Requirements:

Data collection: The card sort procedure can take up to an hour to perform. The SWAT event ratings can be administered during (real time), after (retrospectively), or before (prospectively) the operator performs the task of interest. The operator ratings can be obtained verbally, by paper and pencil, or electronically via a keypad.

Operator training: Practice is needed for the operators to become familiar with the operational definitions and the giving of ratings.

Operator Acceptance: SWAT has been used successfully in aviation and other application. However, cooperation and motivation is the key to obtaining a valid card sort which are the most difficult aspect of this technique.

Safety: Plans must be made as to what to do if the operator is too busy to give real-time ratings. Real-time ratings should be secondary to the primary concern with operational safety (e.g., flying a plane or controlling a land vehicle).

Relative Cost of Use:

Testing time: Card sort can take up to an hour, while the event ratings can be obtained very quickly.

Equipment: Whatever equipment is chosen for data collection. Computer access is necessary for data reduction and analysis.

Setup and support: Careful administration is required, particularly for card sort.

Data analysis: Descriptive and inferential statistics can be used. Parametric statistics are appropriate since conjoint scaling provides an interval scale and they have been used to examine significant differences between mission segments or task variables.

Comments: When used retrospectively, after a long delay, the operator should be aided in recreating the experiences associated with the task when it was previously performed; audio and video recordings of task performance are helpful in this regard. When used prospectively, the operator or subject matter expert should be aided in creating a useful representation of the task as well as the system and operating environment which form the context of the task that is to be rated. In this latter case, the ratings of workload are made to descriptions of tasks and events that have not yet been personally experienced by the individual making the ratings (see Eggleston & Quinn, 1984).

References:

Armstrong Aerospace Medical Research Laboratory (1987, June). Subjective workload assessment technique (SWAT): A user's guide. Dayton, OH: AAMRL, Wright Patterson AFB.

Eggleston, R.G., & Quinn, T.J. (1984). A preliminary evaluation of a projective workload assessment procedure. Proceedings of the Human Factors Society 28nd Annual Meeting (pp. 695-699). Santa Monica, CA: Human Factors Society.

Reid, G. B., Eggemeier, F., & Nygren, T. (1982). An individual differences approach to SWAT scale development. Proceedings of the Human Factors Society 26th Annual Meeting (pp. 639-642). Santa Monica, CA: Human Factors Society.

Reid, G. B., Shingledecker, C. A., & Eggemeier, F. T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting (pp. 522-525). Santa Monica, CA: Human Factors Society.

Availability:

Human Engineering Division
U.S. Air Force Armstrong Laboratory
Wright-Patterson Air Force Base, Ohio 45433-6573

The SWAT Scale

Task or Mission Segment: _____

Please mark an X in one choice for each of the three areas below that best describes what you believe to be the task or mission segment workload

I. TIME LOAD

[1] Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.

[2] Occasionally have spare time. Interruptions or overlap among activities occur frequently.

[3] Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.

II. MENTAL EFFORT

[1] Very little conscious mental effort or concentration required. Activity is almost automatic requiring little or no attention.

[2] Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required.

[3] Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.

III. PSYCHOLOGICAL STRESS

[1] Little confusion, frustration or anxiety exists and can be easily accommodated.

[2] Moderate stress due to confusion, frustration or anxiety. Noticeably adds to workload. Significant compensation is required to maintain adequate performance.

[3] High to very intense stress due to confusion, frustration or anxiety. High to extreme determination and self-control required.

TASK ANALYSIS/WORKLOAD (TAWL)

Description: For a given crewmember and scenario, the Task Analysis/Workload (TAWL; Bierbaum, Fulford, and Hamilton, 1990; Hamilton, Bierbaum, and Fulford, 1991) methodology predicts operator overload using a data base of information produced from a task and workload analysis (see TIS on the predecessor McCracken- Aldrich model). Using a top-down approach, a mission is broken down into phases, phases into segments, segments into functions, and functions into tasks. For example, in an AH-64 evaluation (Szabo & Bierbaum, 1986), seven mission phases, 49 segments, 153 functions, and 653 tasks were identified. For the task analysis, the duration of each task is specified as well as the associated crewmember and subsystem. For the workload analysis, a subject matter expert assigns workload ratings (on a scale from 1 to 7) to the auditory, visual, visual-aided, kinesthetic, cognitive, and psychomotor channels for each task. A scenario is defined using segment and function rules. Segment rules specify what functions will be performed sequentially and concurrently by each crewmember within a specific segment. Similarly, function rules specify what tasks will be performed sequentially and concurrently by each crewmember within a specific function. Randomly-occurring tasks are also defined. A scenario timeline is then generated using the segment and function rules. Independent channel workload is estimated for each time snapshot.

Sensitivity: Operator workload at the task level. Can also identify subsystems associated with high workload.

Diagnosticity: Determine how workload varies across time, crew members, channel components (e.g., cognitive, psychomotor), and subsystems.

Inputs: Detailed task analysis defining the low-level task activities required for a mission including task times. Workload ratings for auditory, visual, visual-aided, kinesthetic, cognitive, and psychomotor channels on a scale of 1 to 7 for each low-level task activity. Scenario decision rules indicating the activities to be performed by each operator.

Outputs: Generates a timeline of low-level activities and predictions of workload at fixed half-second intervals and summary reports of workload statistics, overloads, subsystem use, and subsystem impact on the workload of up to four crew members.

Relative Cost of Use:

Testing time: 6 months to develop a baseline model

Equipment: Perkin-Elmer for original TAWL software; IBM-PC compatible for the microcomputer implementation known as TAWL Operator Simulation System (TOSS; Hamilton, Bierbaum, and Fulford, 1991; Fulford, and Hamilton; and Bierbaum, 1990).

Setup and support: Minimal

Data analysis: Minimal

Comments: TAWL has primarily been applied to predict the impact of system design upgrades on workload in Army aviation settings. Recent applications include various Army ground-based crew stations. Computer implementation of this methodology is necessary. The original TAWL software was developed on a Perkin-Elmer minicomputer. The TAWL Operator Simulation System (TOSS) is a microcomputer implementation of the methodology that employs a menu-driven user-computer interface (Bierbaum, Fulford, and Hamilton; 1989). MicroSaint can also be used to implement the methodology.

References:

Bierbaum, C.R., Fulford, L.A., & Hamilton, D.B. (1990). Task analysis/workload (TAWL) user's guide - Version 3.0 (Research Product 90-15). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD S221 865)

Fulford, L.A., Hamilton, D.B., & Bierbaum, C. R. (1990). TAWL operator simulation system (TOSS) Version 4.0. Proceedings of the Human Factors Society 34th Annual Meeting (p. 1096). Santa Monica, CA: Human Factors Society.

Hamilton, D.B., Bierbaum, C.R., & Fulford, L.A. (1991). Task analysis/workload (TAWL) user's guide - Version 4.0 (Technical Report ASI690-330-90). Fort Rucker, AL: Anacapa Sciences, Inc.

Hamilton, D.B., Bierbaum, C. R., & Fulford, L.A. (1991). Task analysis/workload (TAWL): A methodology for predicting operator workload. Proceedings of the Human Factors Society 35th Annual Meeting (pp. 1117-1121). Santa Monica, CA: Human Factors Society.

Szabo, S. M., & Bierbaum, C. R. (1986). A comprehensive task analysis of the AH-64 mission with crew workload estimates and preliminary decision rules for developing an AH-64 workload prediction model, Vol. I. (ASI678-204-86[B]). Ft. Rucker, AL: Anacapa Sciences, Inc.

Availability:

Chief
Army Research Institute
Aviation Research and Development Activity
Attn: PERI-IR (Mr. C. A. Gainer)
Ft. Rucker, AL 36362-5354

TASK LOAD INDEX (TLX)

Description: The TLX is a multidimensional scale that uses an individual weighting procedure to reduce between-subject variability. It was derived from the NASA-Bipolar scales. It is comprised of two procedures: 1) six rating scales covering different dimensions of workload used to rate OWL; and 2) the "Sources of Workload Evaluation" using paired comparisons of the six dimensions to obtain individual weightings of the dimension importance to workload for any task. The ratings and weightings are combined to produce a global workload rating between 0 and 100.

Sensitivity: Has been demonstrated to be sensitive to differences in task loading in a number of different types of tasks.

Diagnosticity: NASA-TLX gives a global rating of workload. However, the six subscales can potentially be examined individually and used for diagnostic purposes.

Intrusiveness: Little, although it does require a judgment. There was concern (as with most subjective measures) that the judgment might interfere with flight duties, but ratings were obtained real-time.

Implementation Requirements:

Data collection: A "Sources of Workload Evaluation" is obtained for each task under study. The procedure uses only 15 paired comparisons and does not require much time to accomplish. The six TLX scales used to obtain ratings can be administered during (real time), after (retrospectively), or before (prospectively) the operator performs the task of interest. The operator ratings can be obtained verbally, by paper and pencil, or electronically via a keypad. It has been suggested that an alternative to collecting "Sources of Workload Evaluation" is to use Raw TLX (i.e., non-weighted TLX scores) (Byers, Bittner and Hill, 1989).

Operator training: Some practice in using and understanding the operational descriptions of the scales would be helpful.

Operator Acceptance: Has been used successfully in real-time and post-flight aviation applications.

Safety: Plans must be made as to what to do if the operator is too busy to give real-time ratings. Real-time ratings should be secondary to the primary concern with operational safety (e.g., flying a plane or controlling a land vehicle).

Relative Cost of Use:

Testing time: The "Sources of Workload Evaluation" takes on the order of 10 minutes to make paired comparisons. The six ratings would not take significant time if the operators were familiar with the scale descriptions.

Equipment: Can be obtained via paper and pencil, or via computer. Video recording equipment is necessary in order to tape operator activity for use in post-test visual recreation

Setup and support: Minimal.

Data analysis: The weighting and global measure computation can be done by hand, although a computer would be helpful. Descriptive and inferential statistics can be applied. Parametric and non-parametric statistics have been used to examine significant differences between mission segments or task variables.

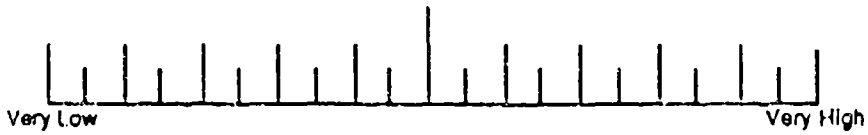
Comments: When used retrospectively, after a long delay, the operator should be aided in recreating the experiences associated with the task when it was previously performed; audio and video recordings of task performance are helpful in this regard. When used prospectively, the operator or subject matter expert should be aided in creating a useful representation of the task as well as the system and operating environment which form the context of the task that is to be rated. In this latter case, the ratings of workload are made to descriptions of tasks and events that have not yet been personally experienced by the individual making the ratings (see Eggleston & Quinn, 1984).

The TLX Scale

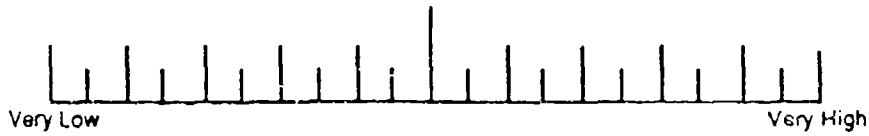
Task or Mission Segment: _____

Please rate the task or mission segment by putting a mark on each of the six scales at the point which matches your experience.

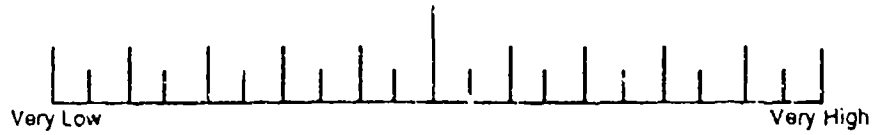
Mental Demand



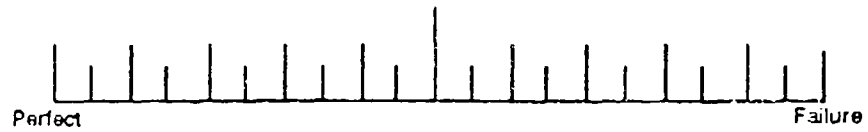
Physical Demand



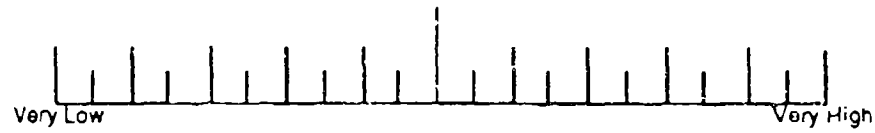
Temporal Demand



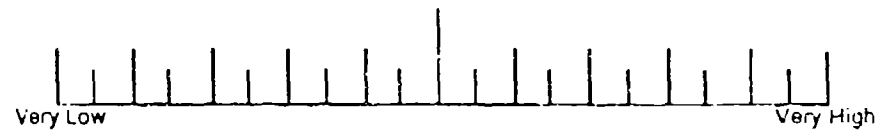
Performance



Effort



Frustration



References:

Eggleston, R.G., & Quinn, T.J. (1984). A preliminary evaluation of a projective workload assessment procedure. Proceedings of the Human Factors Society 28th Annual Meeting (pp. 695-699). Santa Monica, CA: Human Factors Society.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock, & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

NASA-Ames Research Center, Human Performance Group (1986, Feb). Collecting NASA workload ratings: A paper-and-pencil package (Version 2.1). Moffet Field, CA: NASA-Ames Research Center.

Byers, J.C., Bittner, A.C., Jr. and Hill, S.G. (1989). Traditional and raw Task Load Index (TLX) correlations: Are paired comparisons necessary? Advances in Industrial Ergonomics and Safety: Vol. 1. London: Taylor and Francis.

AVAILABILITY:
Human Factors Branch
National Aeronautics and Space Administration
Ames Research Center
Moffet Field, CA 94035

APPENDIX B

WORKLOAD ASSESSMENT OF A MOBILE AIR DEFENSE SYSTEM*

Susan G. Hill Allen L. Zaklad Alvah C. Bittner, Jr.
James C. Byers Richard E. Christ

Abstract

Four operator workload (OWL) scales were retrospectively applied to crew members of a mobile air defense system, the line-of-sight-forward-heavy or LOS-F-H, following a candidate-selection field evaluation: Task Load Index (TLX), Subjective Workload Assessment Technique (SWAT), Overall Workload (OW), and Modified Cooper-Harper (MCH). Jackknife factor analysis revealed the presence of only a single factor (explaining 79.6% of the total variance) and indicated a significant ($p < 0.0075$) ordering of the mean factor loadings: TLX (.935) and OW (.927) were significantly greater than MCH (.862) and SWAT (.860). Multiple correlation also revealed a significant ($p < 0.0001$) relationship, $R = 0.66$, between system performance and TLX ratings. These findings and lessons learned are discussed in the context of the development and validation of a methodology for assessing workload.

INTRODUCTION

Four operator workload (OWL) scales were retrospectively applied to operators of a mobile air defense missile system which was selected subsequent to a recent non-developmental item candidate evaluation (NDICE) field test. This air defense system, the Line of Sight-Forward-Heavy (LOS-F-H), has a primary requirement to engage low-altitude helicopters and fixed-wing threat aircraft as part of the Forward Area Air Defense System. The NDICE was conducted to select a "baseline" LOS-F-H from among four off-the-shelf candidates provided by various teams of contractors. In part, the sensitive nature of the candidate evaluation was responsible for the delay in obtaining access to the cognizant LOS-F-H operators and subject matter experts. As a supplement to the NDICE, the present investigation focused on retrospective assessments of OWL associated with the selected candidate.

Background

A field test to support a non-developmental item candidate evaluation (NDICE) was conducted at Fort Bliss in the late fall of 1987. Four off-the-shelf systems were each use by contractor-trained crews in simulated air defense missions. The simulations consisted of the detection, identification as friend or foe (IFF), and engagement of fixed- and rotary-wing aircraft. Although engagement and firing actions were performed, no live missiles were launched by the crews. During the simulated missions, the crew members participated in no external communications (except to begin and end each mission), and no automatic IFF or command, control and intelligence (C2I) information was provided to the crews.

A total of 25-30 missions were performed by each candidate system under varied test conditions (e.g., conditions of day and night operations). Each mission, lasting about one hour, was composed of four instances or vignettes containing a prescribed number of scripted passes of fixed- and rotary-winged aircraft. The same four vignettes were always used, but they were presented in different random orders throughout the NDICE. Video recordings were made of the actions of the crew members of each candidate system during each mission. Subsequent to the mission, time-locked video monitors provided independent views of each crew member's primary displays and control panels.

Purpose

The objectives of the present investigation were to: (a) explore the applicability of the OWL

* This appendix contains a revised and condensed version of a paper presented at and published in the Proceedings of (pp. 1068-1072) the 32nd Annual Meeting of the Human Factors Society.

scales for obtaining workload assessments 10 weeks subsequent to an operational field test, and (b) evaluate the relationship between system performance and the retrospective workload assessments of the crew members of the selected candidate system.

METHOD

Subjects

The subjects were six soldiers who had been operators of the LOS-F-H during the NDICE. The operators included one radar operator (RO) and five electro-optical operators (EOs). The EOs were junior service members (Private First Class and Specialist 4th Class) and the RO was a Sergeant.

Instruments and Procedures

Prior to the start of the data collection effort a two-hour initial briefing was held with all six subjects to introduce the workload assessment program and the four workload assessment techniques which were to be evaluated. The family of workload assessment scales included: (a) Task Load Index (TLX) (Hart & Staveland, 1987), (b) Subjective Workload Assessment Technique (SWAT) (Reid, Shingledecker, & Eggemeier, 1981), (c) Overall Workload (OW) (Vidulich & Tsang, 1987), and (d) Modified Cooper-Harper (MCH) (Wierwille & Casali, 1983).

Subsequent to the initial group meeting, each operator made workload judgments in conjunction with a review of videotapes (with sound) of his own performance during two specified vignettes in a mission in which he had been a participant. Since we wished to obtain data for comparison purposes, it was decided that we would attempt to get ratings for an "average" mission, one in which the operators were exposed to approximately the same types of mission- and environment-imposed task demands. The mission selected was the same for all operators and was characterized by conditions such as daylight operations, no chemical threat, no obscurant to visual performance, and in the middle to end of the NDICE field test.

Order of video segments was consistent for each operator: (a) an entire mission vignette lasting about 15 minutes was shown and ratings for the overall vignette were obtained; (b) two specific tape segments each showing a different type of attack sequence were shown (one at a time) and ratings obtained; (c) a second vignette was shown and overall ratings were obtained; and (d) a specific segment showing the third type of attack sequence was reviewed and a rating obtained. These individual sessions lasted about 1.5 - 2.0 hours.

After all six subjects had individually viewed tapes and made ratings, they gathered as a group for a final session in which they made workload ratings for the entire NDICE field test. They were also asked to fill out a questionnaire about the workload rating scales and answer questions as to whether they felt they were really able to recall their feelings and experience of workload just from viewing the video tapes. The final session took about 45 minutes.

In summary, over two mission vignettes, each subject made workload judgments for three separate types of passes involving, respectively, two fixed-wing, two rotary-wing, and one rotary-wing aircraft. Within each attack sequence, ratings were made of the workload associated with three operator tasks: visual identification (ID)/IFF, target handoff, and target tracking. (For the single RO, target detection was substituted for the EO task of track to intercept.) In addition, each operator made an overall workload judgement for each vignette, and one for the entire NDICE. These twelve operator workload judgments were made using each of four different rating scales, for a total of 48 ratings per operator. The order of using the four rating scales was counterbalanced over judgments and subjects.

A system performance score for each specific rated mission was provided by the NDICE Test Officer. These integer scores were 0, 1, or 2, reflecting the number of rotary-wing or fixed-wing threat aircraft destroyed in a given pass.

RESULTS

Analyses were conducted in three phases which respectively examined: (a) the factor validities of the four workload scales; (b) the relationship between system performance and the retrospective workload assessments; and (c) a summary of other results relevant to the measurement of workload, to include data from the rating scale questionnaire and interview administered during the final group meeting with the subjects.

Factor Validity Analyses

The factor validity analyses were conducted in two stages. During the first stage, Principal Component Analysis (PCA) was conducted on the 72 sets of segment ratings collected across all subjects and missions using BMDP4M (Dixon, 1983). Each set included global workload ratings using four scales: TLX, SWAT, OW, and MCH. (The mean and standard deviation of global workload ratings for each scale are in Data Attachment B-1 at the end of this appendix.) This analysis revealed a single component, hereafter termed the **OWL factor**, which explained 79.6% of the total variance (the second eigenvalue was only 0.42). The results of this initial analysis supported the view that the four workload scales essentially provide assessments of a single common factor. (The factor scores for each subject's set of 12 workload judgments are in Data Attachment B-2.)

Jackknife PCAs were conducted of the workload measures during the second stage in order to evaluate the stability of the factor loadings of the four scales (i.e., correlations with the OWL factor). Jackknife analysis, it is noteworthy, generally involves successive analyses (PCAs in the present case) dropping subjects one-at-a-time from a data set in order to provide for analysis of the stability of parameter estimates (Hinkley, 1983). In the present case with four factor loadings and the 6 subjects, a 4 (loadings) by 6 (subject dropped) matrix was produced which could be analyzed by conventional repeated measures analysis of variance (ANOVA). The ANOVA, using BMDP2V (Dixon, 1983), revealed a very highly significant difference between the workload scale factor loadings ($F(3,15) = 17.05$, $p < .0075$). Subsequent analysis revealed the following ordering of the mean factor loadings:

TLX(.935), OW(.927), MCH(.862), SWAT(.860).

The TLX-OW difference is statistically significant but negligible in practical terms, the MCH-SWAT difference is insignificant, but all other differences are significant.

OWL and Performance Relationships

Two stepwise regression analyses (BMDP2R) were conducted to explore the relationship between system performance and operator workload (Dixon, 1983). In the first analysis, the dependent variable (PERF) was the system performance score and the independent variables included: the TLX rating of global workload; as well as six dichotomous variables which indexed the subject making the rating (ID1-ID6). Stopping after accretion of three variables (TLX, ID4 and ID6), this analysis revealed a substantial multiple correlation, $R = 0.66$ which was very highly significant ($F(3,44) = 11.12$, $p < .0001$). The resulting model for system performance (PERF) was:

$$PERF = 2.069 - 0.013*TLX - 1.077*ID4 + 0.526*ID6 \quad [Eq.1]$$

The ID4 and ID6 weights, in this model, indicate lesser and greater than average performance for a given level of TLX for the respective subjects (4 and 6). However, this model altogether predicts generally decreasing performance (PERF) with increases in workload (TLX) across all subjects.

The second regression analysis reversed the first-analysis' respective independent and dependent variable roles for TLX and PERF in order to establish estimates of TLX for integer levels of PERF (0, 1, and 2). [This, it is noteworthy, was judged to be a more pertinent way to express the TLX-PERF relationship for some analysts.] The dependent variable consequently was TLX and the independent variables included PERF as well as the six dichotomous variables which indexed the subject making the rating (ID1-ID6). Stopping after accretion of three variables (PERF, ID4 and ID6), this analysis not unexpectedly revealed results paralleling those for the first regression analysis ($R = 0.50$, $p < .001$). Figure B-1 illustrates the resulting model where a 0-targets-destroyed value of PERF is associated with a predicted TLX of 59.5 and 2-targets is associated with 29.0 for the "average subject".

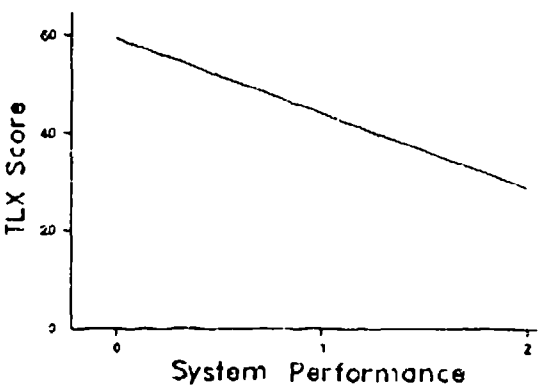


Figure B-1. *The relationship between workload ratings and system performance in the LOS-F-H NDICE study.*

Other Results Relevant to the Measurement of Workload

Two sets of operator performance time data were obtained which reflect on the characteristics of the four rating scales. The first set of time measurements were obtained during the initial group meeting with the subjects; the time was measured for each subject to complete the procedures required to use the two multidimensional scales (see the references cited above for the SWAT and TLX scales). The times it took the six soldiers to complete the SWAT card sort procedure were 25, 30, 33, 34, 43, and 45 minutes (mean time = 35 minutes with a standard deviation of 7.7 minutes). The times it took soldiers to complete the TLX paired comparison procedure were approximately 5-7 minutes for the first task to which the procedure was applied and 2-3 minutes for subsequent comparisons.

The second set of measurements was a sample of approximate times to complete the four rating scale techniques during the meetings at which individual soldiers rated their video taped mission. Table B-1 gives the means and standard deviations of these scale completion times along with respective sample sizes. It may be seen in Table B-1 that it required considerably less time to complete the OW scale than any of the other three scales; more time was required to complete the TLX weighted rating scales than the SWAT or MCH scales.

...ads) to Complete Workload Rating Scales

Study	n	Mean	SD
TLX	38	51.3	29.5
OW	33	9.8	8.4
MCH	27	29.1	26.3
SWAT	27	33.6	24.6

Table B-2 shows the frequency of times each scale was ranked first according to general preference (i.e., being liked), being easy and being difficult to complete, and permitting a subject to express his workload experiences. It may be seen that a majority of the subjects preferred either the TLX or the OW over the other two scales. Almost all subjects agreed that the OW scale was the easiest to complete but they divided almost equally

Table B-2

Operator Acceptance of Workload Rating Scales in the LOS-F-H NDICE Study

Rating Scale			
TLX	OW	MCH	SWAT
Which of the questionnaires did you like the best?			
2	2	1	1
Which questionnaire was the easiest to fill out?			
1	4	1	0
Which questionnaire was the hardest to fill out?			
0	1	3	2
Which questionnaire do you think best allowed you to describe the workload you experienced?			
5	0	1	0

Note. Data shown are the number of times each scale is given the highest ranking.

in indicating that MCH and SWAT were the most difficult. All but one subject indicated that the TLX technique best allowed them to describe their workload experiences.

An analysis of the data from the SWAT card sorts revealed some problems with this procedure. Out of six subjects, four did not have truly acceptable sorts (according to the SWAT User's Guide, AAMRL, 1987). This problem arose due to excessive violations of the axioms which underlie the mathematical model used to derive workload scores from the operator ratings.

The questionnaire and interviews also asked the subjects to indicate the extent to which they were really able to recall their feelings and experience of workload just from viewing the video tapes. Five conclusions may be derived from these recall data:

- Some soldiers could, some were less sure that they could reliably recall workload experiences by looking at video tapes of themselves during missions that had been performed more than three months earlier.

- Unless something unusual happened, some operators seemed to have a difficult time differentiating a particular mission segment from others of the same kind. They seemed to view a mission segment (e.g., two-fixed wing aircraft) and give it a rating for the generic case rather than the specific case that was captured on the video recording.

- There seemed to be some difficulty in differentiating tasks within a short duration segment (e.g., when the detection task ends and the identify task begins).

- There seemed to be some difficulty differentiating performance from other factors of workload. For some of the operators, if they felt they had performed poorly in a video tape segment they had just viewed, they would rate workload high, even if they also indicated that the particular task in question was neither difficult or excessively demanding.

- The missions which were actually conducted during a field test can be substantially different from the ones which were planned and programmed to have occurred. This, in turn, made mission vignettes which were supposed to be the same over all test missions different from each other. Consequently, although there was an attempt to use video recordings of the same mission vignettes for all operators, there were substantial differences in the vignettes.

DISCUSSION

This investigation evaluated the retrospective use of four OWL assessment scales following a candidate selection field test and explored the relationships between system performance and workload as measured by one of those scales (i.e., TLX). The results obtained with the four scales are evaluated in this section in terms of their contribution to the development and validation of a methodology for estimating and evaluating OWL in Army systems. The results obtained from relating workload and system performance are discussed in terms of the potential usefulness of OWL measures.

Retrospective Application of OWL Scales for Field Tests

This investigation demonstrated the successful retrospective application of a family of OWL measures 10 weeks subsequent to a field test. This work was consequently performed under constraints that are more severe than most previous applications of such scales, but are not uncommon in many tests and evaluations of Army systems. The use of mission video tapes, it is believed, facilitated the retrospective application of the OWL scales as most (but not all) soldier-operators felt comfortable recalling workload after the 10 week hiatus.

No doubt, more detailed mission-specific information could have been obtained under more desirable assessment conditions. For example, it would have been desirable for the OWL data collection team to participate in test planning and to have made real-time observations of test performance to guide subsequent assessment and interpretation of OWL. Such information would have provided for timely study of specific problems and events (i.e., as they occurred). However, the present application of OWL measures yielded formal and informal guidance regarding the retrospective use of OWL scales under field conditions.

Formal guidance. The four OWL measurement scales were shown to have clearly different factor validities in this investigation. The TLX scale had the greatest and the MCH and SWAT scales had the least factor validities in this investigation; OW was statistically different from each of the other three though not practically different from TLX. The rating scale questionnaire results shown in Table B-2 indicate that most subjects thought that TLX was one of the easiest to complete and the best scale for describing their workload experiences. On the basis of all these results, one could be tempted to solely recommend TLX.

However, as seen in Table B-1, TLX

individual assessments required more time-to-complete than the other measures. Except for the more than 5-fold time-to-complete of TLX relative to OW, these completion-time differences may be judged relatively marginal in the context of other time costs (such as the mission video assessments that were employed here). Consequently, given the high factor validity of OW and its generally favorable ratings in the questionnaire, arguments may be made for its use for screening very large numbers of mission segments and operator tasks with respect to overall workload (e.g., in preparation for more diagnostic evaluation of "workload problem areas"). These arguments, it is noteworthy, are predicated on tradeoffs of temporal cost, scale validity, and subject availability factors which may be evaluated only on a case-by-case basis.

In summary, the results of the present investigation point toward use of TLX, because of its consistently high factor validity, for all but screening applications. In the latter case it may be more appropriate to use OW.

Informal guidance. Experience administrating the OWL scales during the present investigation point toward three sets of informal guidance for future application of OWL measurement scales:

- The initial briefing, separate from the mission data collection, serves as a convenient time to introduce the data collection team, the concept of workload, and the workload ratings tools. The procedures required to use the multidimensional SWAT and TLX scales may also be obtained at this time. This initial briefing did entail coordination to ensure the presence of all potential subjects.

- The required SWAT sorts may not be satisfactorily accomplished by all subjects. In the present investigation, 4 out of the 6 operators had excessive axiom violations according to the SWAT User's Guide. Consequently, time must be set aside for potentially resolving such problems (we have encountered subjects where this has proven not possible). Hence, the experimenter must also be prepared to either use subjects despite such inconsistencies or discard them.

- The importance of talking with the crews to obtain their impressions of "what they do and why" was confirmed during this test. Informal discussions with crews give added insight into potential workload and other human factors problems.

Relationship of Performance and Workload

The substantial and highly significant multiple correlations between measures of system performance and workload (R = .50 and .66) were consistent with theoretical expectations. In particular, the model derived from the regression of system performance onto workload (Eq. 1) indicates generally decreasing performance (PERF) with increases in workload (TLX). Of interest, modulating this relationship were individual differences indicating lesser and greater than average performance for a given level of TLX (for Subjects 4 and 6, respectively). Such differences, it is noteworthy, could arise because: (a) the performance of some operators is more, or less sensitive, to a given workload level than for typical subjects (perhaps reflecting cognitive strategies or personality difference; or (b) OWL reports of some subjects reflect relative over- or under-statements of experienced workload (reflecting personal biases in reporting). Unfortunately, neither of these possibilities may be resolved from the results of the present investigation, but remain open questions for future research in other contexts.

This investigation, it may be recalled, was aimed at exploring the applicability of the OWL scales for obtaining retrospective workload assessments after a delay of several weeks. The substantial and highly significant multiple correlations between system performance (PERF) and workload (TLX) shown in this investigation support the efficacy of such an application.

CONCLUSIONS

Two broad conclusions can be drawn from the present evaluation of the use of the OWL scales under field test conditions.

(1) TLX consistently had the highest validity in the present field test and may be recommended for all but screening applications where it may be appropriate to use OW.

(2) Operator workload (OWL) measures may be applied and evaluated in the stringent retrospective environments which characterize many Army test and evaluation efforts.

REFERENCES

Armstrong Aeromedical Research Laboratory (AAMRL). (1987). Subjective workload assessment technique (SWAT): A user's guide. Dayton, OH: AAMRL, Wright-Patterson Air Force Base.

Dixon, W.J. (Ed.). (1983). BMDP statistical software. Los Angeles, CA: University of California Press.

Hart, S.G. & Staveland, L.E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.S. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Hinkley, D.V. (1983). Jackknife methods. In S. Kotz, N.L. Johnson, & C.B. Read (Eds.) Encyclopedia of statistical sciences: Vol.4 (pp. 280-287). New York: Wiley.

Reid, G.B., Shingledecker, C.A., & Eggemeier, T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting (pp. 522-525). Santa Monica, CA: Human Factors Society.

Vidulich, M.A. & Tsang, P.S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

Wierwille, W.W. & Casali, J.G. (1983). A validated rating scale for global mental workload measurement application. Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.

DATA ATTACHMENT B-1

COMPARISON OF WORKLOAD RATING SCALES FOR THE LOS-F-H NDICE STUDY

MISSION CONDITION	TASK SEQUENCE	RATING SCALE TLX	OW	MCH	SWAT
		---------- MEANS ----------			
1 Rotary Wing	Visual ID/IFF	21.66	13.33	9.16	10.00
	Handoff	26.00	34.16	14.66	17.33
	Track/Detect	18.00	22.50	9.16	8.33
2 Rotary Wing	Visual ID/IFF	28.16	18.33	12.83	27.66
	Handoff	42.66	33.33	24.00	30.66
	Track/Detect	40.83	40.00	27.66	54.16
2 Fixed Wing	Visual ID/IFF	37.83	27.50	18.33	36.00
	Handoff	47.66	46.66	29.50	44.33
	Track/Detect	51.83	53.33	31.33	65.33
		---- STANDARD DEVIATIONS ----			
1 Rotary Wing	Visual ID/IFF	17.10	14.02	8.28	13.19
	Handoff	15.00	29.90	13.32	22.71
	Track/Detect	12.56	19.17	8.28	9.30
2 Rotary Wing	Visual ID/IFF	11.77	12.11	8.28	34.37
	Handoff	33.35	26.01	27.70	39.82
	Track/Detect	31.17	28.10	25.15	41.23
2 Fixed Wing	Visual ID/IFF	27.65	27.34	16.56	38.36
	Handoff	22.33	28.22	18.25	34.12
	Track/Detect	23.82	23.80	20.52	40.04

DATA ATTACHMENT B-2

FACTOR SCORES FOR ALL SUBJECTS LOS-F-H NDICE

Commander (RO)	IFF	Handoff	Detect	
2 Rotary Wing	-0.72	-0.72	-0.69	-0.71
2 Fixed Wing	-0.64	-0.49	-0.21	-0.45
1 Rotary Wing	-1.19	-1.18	-0.91	-1.09
	-0.85	-0.80	-0.60	-0.75

Gunner (EO) 1	ID	Handoff	Track	
2 Rotary Wing	-0.68	2.30	1.08	0.90
2 Fixed Wing	1.42	1.19	1.88	1.50
1 Rotary Wing	-1.16	-0.01	-1.26	-0.81
	-0.14	1.16	0.57	0.53

Gunner (EO) 2	ID	Handoff	Track	
2 Rotary Wing	-0.33	-0.46	-1.38	-0.72
2 Fixed Wing	-1.43	-0.66	0.58	-0.50
1 Rotary Wing	-1.37	-1.30	-1.06	-1.24
	-1.04	-0.81	-0.62	-0.82

Gunner (EO) 3	ID	Handoff	Track	
2 Rotary Wing	-0.98	-1.31	-0.86	-1.05
2 Fixed Wing	-0.50	1.91	-0.40	0.34
1 Rotary Wing	-0.89	-0.51	-0.83	-0.74
	-0.79	0.03	-0.70	-0.49

Gunner (EO) 4	ID	Handoff	Track	
2 Rotary Wing	-0.86	1.01	1.42	0.52
2 Fixed Wing	-0.83	-0.01	0.18	-0.22
1 Rotary Wing	-0.42	1.07	-0.84	-0.06
	-0.70	0.69	0.25	0.08

Gunner (EO) 5	ID	Handoff	Track	
2 Rotary Wing	0.09	-0.42	1.58	0.42
2 Fixed Wing	0.81	0.66	1.72	1.06
1 Rotary Wing	0.09	-0.25	0.36	0.07
	0.33	0.00	1.22	0.52

APPENDIX C

GENERIC WORKLOAD RATINGS OF A MOBILE AIR DEFENSE SYSTEM*

Alvah C. Bittner, Jr. James C. Byers Susan G. Hill
Allen L. Zaklad Richard E. Christ

Abstract

Operator workload (OWL) scales were used to obtain ratings of generic mission scenarios and tasks for a mobile air defense system (the line-of-sight-forward-heavy or LOS-F-H) following a field test in support of a systems evaluation program. Task Load Index (TLX), Subjective Workload Assessment Technique (SWAT), Overall Workload (OW), and Modified Cooper-Harper (MCH) ratings were obtained from both crew members and subject matter experts (SMEs) of the system. Jackknife factor analysis revealed the presence of only a single OWL factor for both operators and SMEs (explaining 75.9% and 82.6% of the respective total variances) and indicated a significant ($p < .00005$) ordering of the mean factor loadings: TLX (0.924) was significantly greater than OW (0.905) and MCH (0.904), both of which were greater than SWAT (0.778). Subsequent analysis of OWL factor scores indicated that the highest levels of OWL were obtained for the track-to-intercept task during rotary-wing and fixed-wing attacks although the identify as friend or foe task during a dual rotary-wing attack was almost as high. These findings are discussed in the context of a methodology for assessing OWL.

INTRODUCTION

Operator workload (OWL) assessments were obtained for a mobile air defense missile system, the Line-of-Sight-Forward-Heavy (LOS-F-H). A previous OWL study of this system (Hill, Zaklad, Bittner, Byers, & Christ, 1988 -- see Appendix B of this report) found that performance and workload were related, but did not find a relationship between OWL ratings and critical mission conditions (e.g., type of attack sequence). It was suggested that the ratings reflected idiosyncratic differences in specific mission segments which washed out the effects of the mission variables. The approach taken in this study to overcome such mission-specific quirks (and the small number of data points) was to collect workload ratings of generic rather than actual missions. This study also explored the differences in OWL ratings between operators (LOS-F-H crew members) and other kinds of subject matter experts (SMEs).

Background

The previous study investigated the **retrospective** application of operator workload scales to LOS-F-H crew members after they had reviewed videotapes of their own performance during an "average" mission. Average missions were ones which presumably exposed the operators to approximately the same types of mission- and environment-imposed task demands. Consequently, variations in OWL ratings should have reflected differences in the workload associated with different mission-specific operator tasks. The results, however, showed that there were large variations in OWL ratings across crew members within the same "average" mission segments; these clouded statistical comparisons of the segments and tasks of interest.

In hind-sight, it seems that the missions which were actually conducted were probability substantially different from the ones which were programmed to have occurred. Since our attempt to use video recordings of an average mission was based on the type of mission which was suppose to have occurred, there is the possibility that there were in fact substantial differences in these missions. If so, the OWL ratings obtained would have reflected idiosyncratic differences in specific mission segments. These differences in missions

---

* This appendix contains a revised and condensed version of a paper presented at and published in the Proceedings of (pp. 1476-1480) the 33nd Annual Meeting of the Human Factors Society.

would have led to large variations across subjects in workload ratings for the same types of mission segments and task.

Purpose

The objectives of this study were to (a) investigate the applicability of workload ratings to generic missions, and (b) compare the workload ratings of experienced system operators and other subject matter experts.

METHOD

Subjects

There were two groups of subjects: LOS-F-H crew members and SMEs. The crew members were five electro-optical operators (EOs) who had been participants in the previous non-developmental item candidate evaluation (NDICE) field test and had participated in the previous OWL data collection effort associated with that test. No radar operators (ROs) were available for the present study. The SMEs were nine civil service and contractor civilians who had been or would be working directly in the LOS-F-H program. They had a diverse range of experience with the system: four were associated with manpower, personnel, and training analyses while the other five were from training organizations. All were associated with supporting U.S. Army organizations and agencies. Table C-1 delineates the experience of the SMEs.

Procedure and Instruments

The workload assessments of the two groups of subjects occurred during two separate data collection sessions. These sessions took place approximately six months subsequent to the NDICE. At the beginning of the sessions the SMEs were introduced to and the crew members reviewed, as necessary, the general objectives of the workload assessment program and the four workload assessment techniques which were to be evaluated.

The rating techniques were: (a) Task Load Index (TLX) (Hart & Staveland, 1987), (b) Subjective Workload Assessment Technique (SWAT) (Reid, Shingledecker, & Eggemeier, 1981), Overall Workload (OW) (Vidulich & Tsang, 1987), and (d) Modified Cooper-Harper (MCH) (Wierwille & Casali, 1983). All subjects were briefed about the specific purpose of their participation in the present study and necessary procedures were completed for using the two multidimensional rating techniques.

Operator workload assessments using each rating technique were made by each subject for nine combinations of three mission conditions and three task segments. The order of using the four rating scales was counterbalanced over judgments and subjects. Mission conditions were a "single rotary-wing (RW) attack"; a "dual RW attack"; and a "dual fixed-wing (FW) attack." Task segments were visual Identification/Identify as Friend or Foe (ID/IFF); Handoff of a target track by the RO to the EO; and Track-to-Intercept. Each individual was given a packet of OWL forms, each form marked with a specific combination of a mission

Table C-1

Experience of SMEs in LOS-F-H Generic Study

SME	ASSOCIATION WITH SYSTEM	INVOLVEMENT IN NDICE	TRAINING ON SYSTEM	WATCHED FILMS OF NDICE (10 OR MORE)	OTHER AIR DEFENSE EXPERIENCE	MILITARY EXPERIENCE
1	MANPRINT	YES	YES	YES	YES	YES
2	MANPRINT	NO	YES	YES	NO	YES
3	MANPRINT	YES	YES	YES	YES	YES
4	MANPRINT	YES	YES	YES	YES	YES
5	TRAINING	YES	YES	YES	YES	YES
6	TRAINING	NO	NO	YES	YES	YES
7	TRAINING	NO	YES	NO	YES	YES
8	TRAINING	NO	YES	NO	YES	NO
9	TRAINING	NO	NO	NO	YES	YES

condition and task segment. After the relevant "generic" mission was defined by the data collector, the subjects were asked to rate the workload associated with that mission condition and task segment over all their relevant experiences with the LOS-F-H system. The SMEs not familiar with the LOS-F(H) system or NDICE were requested to base their ratings on their knowledge of similar systems and tests. The crew members made OWL judgments only for the tasks which they (EOs) perform. The SMEs were asked to made OWL judgments for both RO and EO tasks. All subjects were also asked to make OWL judgments of an "average LOS-F-H mission."

RESULTS

Analyses were conducted in two phases which were directed at (a) comparison of the factor validities of the four workload scales as rated by crew members and SMEs; and (b) evaluation of crew member and SME workload variations across generic mission conditions and task segments.

Factor Validity Analyses

The factor validity analyses were conducted in two stages. During the first stage, Principal Components Analyses (PCAs) were separately conducted on the respective complete sets of 50 crew member and 80 SME mission segment ratings using BMDP4M (Dixon, 1983). For both groups, each complete set included global workload ratings using four scales: TLX, SWAT, OW, and MCH. (The means and standard deviation of global workload ratings for each scale are in Data Attachment C-1 at the end of this appendix.) Data from 5 SMEs, as will be discussed later, could not be used because of problematic MCH or SWAT ratings. The PCA analyses both revealed single components which respectively explained 75.9% and 82.6% of the crew member and SME total variances (the second eigenvalues were only 0.57 and 0.40). The results of this initial stage of analysis suggested that for both groups the four workload scales essentially assess a single common OWL factor. (The factor scores for each subject's workload judgments are in Data Attachment C-2.)

Jackknife PCAs were separately conducted of the crew member and SME OWL ratings data sets during the second stage of analysis to provide the basis for comparing group OWL factor loadings. Jackknife analysis, it is noteworthy, generally involves successive analyses (PCAs in the present case) dropping subjects one-at-a-time from data sets in order to provide for analysis of the stability of parameter estimates (Hinkley, 1983). In the present case, the crew member Jackknife PCAs resulted in a 4 (loadings) by 5 (subject-dropped) matrix. The SME Jackknife PCAs resulted in a 4 (loadings) by 4 (subject-dropped) matrix. Treating these two matrices as grouped repeated measures data, an analysis of variance (ANOVA) may be used to evaluate group and OWL scale loading differences. Using BMDP2V (Dixon, 1983), ANOVA revealed a very highly significant difference between the workload scale factor loadings ($F(3,21) = 25.12$, Huynh-Feldt $p < 0.00005$). Subsequent analysis revealed the following ordering of the mean factor loadings:

TLX(.924), OW(.905), MCH(.904), SWAT(.778),

where, excepting OW-MCH, all differences were statistically significant ($p < 0.05$). The interaction of scale and group (SxG) was also found significant ($F(3,21) = 8.25$, Huynh-Feldt $p < 0.005$), although the overall difference between the grand mean of all ratings for the crew member (0.857) and SME (0.903) groups was nonsignificant ($F(1,7) = 2.30$, $p > 0.17$). Explaining less than a third of the variance as the scale main effect, the SxG interaction was attributable to differences in the SWAT and MCH factor loadings for the two groups. Interestingly, the SWAT ratings substantially differed although both represented the minimum loadings for their respective groups [crew member (0.719) vs. SME (0.851)]. The difference in the group MCH loadings was substantially less (0.037) and appeared less interesting [because of problems experienced by the excluded SMEs in properly using the instrument]. Supporting this interpretation, the residual SxG interaction was found nonsignificant after eliminating group differences in SWAT and MCH ($F(1,21) = 2.97$, $p > 0.09$). The results altogether essentially support the ordering of the mean factor loadings.

Workload Analyses

An ANOVA was conducted to examine the effects of LOS-F-H system variables on operator workload as assessed by OWL factor scores. BMDP4M (Dixon, 1983) was first used to develop the OWL factor scores as an output from a PCA of data from the five crewmembers and, after dropping two who did not properly perform the MCH ratings, seven of the SMEs. Repeated measures ANOVA

using BMDP2V (Dixon, 1983) was then used to evaluate the effects of Group (crew member vs. SME), Mission Condition (single RW, dual RW, and dual FW), and Task Segment (ID/IFF, handoff, and track-to-intercept). Of greatest relevance to the question of using SMEs versus crewmembers to evaluate OWL, this ANOVA found that neither the Group main effects ($p > 0.78$) nor any of the interactions of group and the other variables were significant ($p > 0.12$). This indicates that LOS-F-H crew members and SMEs yield equivalent evaluations of operator workload over the system variables investigated.

The ANOVA of the OWL factor scores also revealed significant effects for Mission Condition ($F(2,20) = 5.76$, Huynh-Feldt $p < 0.011$), Task Segment ($F(2,20) = 3.74$, Huynh-Feldt $p < 0.05$), as well as the interaction of Mission Condition and Task Segment ($F(4,40) = 2.54$, Huynh-Feldt $p = 0.05$). Figure C-1 illustrates the nature of these main and interaction effects.

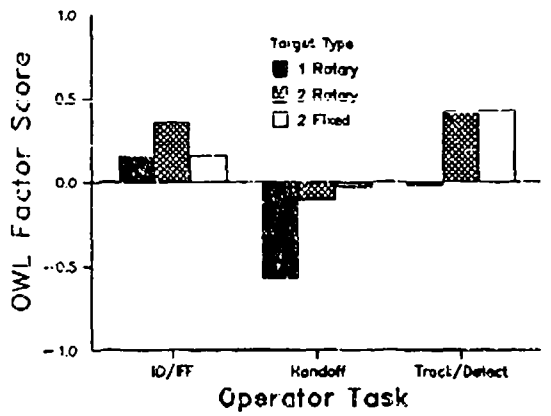


Figure C-1. *The effect of operator task and target type on workload in the LOS-F-H.*

Examining this figure, it may be seen that the mean single RW OWL factor score (-0.15) is substantially less than those for dual RW (0.22) or single FW (0.19). It may likewise be seen that the mean handoff factor score (-0.24) is substantially less than those for ID/IFF (0.23) or track-to-intercept (0.28). Lastly, the nature of the mission condition-task segment interaction may be seen. Namely, ID/IFF during the dual RW mission condition (0.36) is substantially greater than that for the dual FW and single RW conditions which are essentially equal (0.17 vs. 0.16). However, for the handoff and track events, the two dual mission conditions resulted in essentially equal mean OWL factor scores while that for single RW was at a substantially lower level. These results altogether indicate that the highest levels of OWL were obtained for track-to-intercept during dual RW and FW attacks with ID/IFF during a dual RW attack almost as high.

Analysis of TLX Subscales

Due to limitations in time, a limited examination was made of the ratings obtained from the five crew members for each of the six TLX subscales. This cursory analysis showed that there was a significant difference in the ratings obtained from the subscales, $F(5,20) = 5.47$, $p < .01$. In order of decreasing magnitude the mean weighted subscale scores are: Temporal Demand (56), Mental Demand (40), Performance (32), Effort (29), Frustration (14), and Physical Demand (2). Separate analyses performed for each subscale showed no significant variation in any due to mission conditions or task segments.

DISCUSSION

This investigation evaluated the use of four OWL scales to obtain workload ratings of both experienced system operators and other SMEs for generic missions of the LOS-F-H system. The discussion which follows addresses (a) the efficacy of the OWL scales for these two groups of raters, (b) the usefulness of generic mission descriptions for evaluating workload effects, and (c) the implications of the workload results obtained for the system under study.

OWL Assessments From Operators and SMEs

This investigation demonstrated the successful application of the OWL scales for workload evaluations by operators and some SMEs. Not all SMEs, as noted earlier, could be used in the analyses because of a variety of problems. In particular, four of the nine SMEs did not produce acceptable SWAT sorts and two of the nine did not follow procedure for completing MCH scales (with one overlap). Consequently, a total of five SMEs were excluded from the factor validity analysis, and the two who had difficulty with MCH were necessarily excluded from the workload analyses. Interestingly, the Table C-1 experience variables appeared to be unrelated to the SWAT and MCH difficulties experienced by some SMEs. The equivalence of operators and SMEs is discussed in terms of both the OWL factor validity and the

LOS-F(H) workload analyses in the remainder of this section.

The OWL factor validity analysis revealed a very highly significant main effect difference between the workload scales ($p < 0.00005$). Although there was some evidence of a group-by-scale interaction in the factor validity analysis ($p < 0.005$), the result also indicated that the two groups had equivalent orderings for the two measures with the highest validities: TLX (0.924) and OW (0.905). These results, it is pertinent to observe, support our previous recommendations of TLX for precision applications and OW for screening purposes (Hill et al..1988). The OW scale may again be recommended for screening because it continues to exhibit modest but consistent OWL factor validities while requiring substantially less time-to-complete (20% of TLX as shown by Hill et al., 1988). The TLX scale again may be recommended for precision evaluations because it continues to manifest significantly greater factor validities than the other scales (cf., Byers, Bittner, Hill, Zaklad, & Christ, 1988; Hill et al., 1988 - Appendices G and B, respectively).

Operators and SMEs were found also essentially equivalent in terms of their OWL factor scores across evaluated conditions. Although there were significant Mission Condition and Task Segment effects, neither the main effects of group ($p > 0.78$) nor any of its interactions with these other variables were significant ($p > 0.12$). These results suggest that SMEs may be expected to give essentially equivalent results to operators in evaluations similar to the present (provided they acceptably use the scales).

Workload Ratings of Generic Mission Ratings

Generic ratings proved useful for minimizing idiosyncratic mission differences. As described earlier, analysis revealed significant effects for Mission Condition, Task Segment, and their interaction. This wealth of significant findings using generic ratings stands in sharp contrast to the earlier found paucity with specific ratings (Hill et al., 1988). Of course, means of very much larger numbers of specific ratings also could be expected to yield a similar wealth of results. Such means certainly would appear to be preferred in terms of having higher face validity. However, the temporal and other costs of obtaining sufficient numbers might well be prohibitive in the context of many investigations (e.g., Hill et al., 1988). In addition, SMEs may be the only available source of ratings as access to operators can be extremely limited or impossible. Representing "subject averages" across missions, ratings of generic missions consequently appear more widely applicable for overcoming idiosyncrasies than increasing sample sizes. Generic ratings should be considered for application where either only a small number of missions can be rated or the only practicable operator workload raters are SMEs.

Impact of Workload for the LOS-F-H System

Analysis of the OWL factor scores revealed a significant interaction of missions and segment which was illustrated in Figure C-1. As was seen in this figure, the highest levels of OWL were obtained for: ID/IFF during an attack by dual RW; and track-to-intercept during attacks by either dual RW or dual FW. The high level for ID/IFF during a dual RW attack was not unexpected as there was typically little time to identify both RWs which pop-up relatively close to the fire unit and pose substantial threat. The cursory analysis of TLX subscales showed, not surprisingly, that the global rating had a large temporal demand component. Workloads associated with ID/IFF and track-to-intercept it may be noted, would be expected to be significantly reduced with implementation of an automatic system for ID/IFF. These results point toward both the nature of the highest workload conditions and possible means for reduction.

CONCLUSIONS

Three broad conclusions may be drawn from the present evaluation of the use of OWL scales:

(1) Generic ratings may be used to assess mission conditions and task segments while minimizing differences caused by specific mission idiosyncrasies. These should be considered for application when either only a small number of missions can be rated or only SMEs are available.

(2) There were no systematic differences found between generic OWL ratings made by SMEs and crew members who had operated the system. This suggests that SMEs, who do not necessarily have specific experience with the system of concern, can still provide meaningful quantitative OWL information for generic missions when crew

members are not available.

(3) It would be a mistake to assume that anyone called an SME could make equivalent OWL judgments to experienced system operators. SMEs should be used with caution to evaluate generic operator workload pending a more complete understanding of needed rater characteristics for judgment of operator workload.

REFERENCES

Byers, J. C., Bittner, A. C., Jr., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1145-1149). Santa Monica, CA: Human Factors Society.

Dixon, W. J. (Ed.). (1983). BMDP statistical software. Los Angeles, CA: University of California Press.

Hart, S. G., & Staveland, L. E. (1987). Development of a NASA TLX (Task Load Index): Results of empirical and theoretical research. In P.S. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Hill, S. G., Zaklad, A. L., Bittner, A. C., Jr., Byers, J. C., & Christ, R. E. (1988). Workload assessment of a mobile air defense missile system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1068-1072). Santa Monica, CA: Human Factors Society.

Hinkley, D. V. (1983). Jackknife methods. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), Encyclopedia of statistical sciences: Vol. 4 (pp. 280-287). New York: Wiley.

Reid, G. B., Shingledecker, C. A., & Eggemeier, T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting (pp. 522-525). Santa Monica, CA: Human Factors Society.

Vidulich, M. A., & Tsang, P. S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement application. Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.

DATA ATTACHMENT C-1

COMPARISON OF WORKLOAD RATING SCALES LOS-F-H GENERIC

Mission/Task	OW	MCH	SWAT	TLX
	---------- MEANS ----------			
1 Rotary Wing				
Visual ID/IFF	50.41	42.50	60.52	49.05
Handoff	37.08	24.91	43.16	38.22
Track/Detect	47.91	36.00	60.55	44.91
2 Rotary Wing				
Visual ID/IFF	55.00	43.41	68.83	51.47
Handoff	47.08	35.16	52.43	44.74
Track/Detect	57.91	42.50	74.99	50.72
2 Fixed Wing				
Visual ID/IFF	51.66	37.83	66.96	48.88
Handoff	47.08	35.16	53.45	45.00
Track/Detect	56.66	44.33	67.51	50.97
	------- STANDARD DEVIATION ------			
1 Rotary Wing				
Visual ID/IFF	21.45	22.41	25.34	15.38
Handoff	23.36	22.51	27.26	16.50
Track/Detect	23.54	24.46	29.40	17.59
2 Rotary Wing				
Visual ID/IFF	20.97	25.87	28.27	17.19
Handoff	29.47	26.45	37.80	21.97
Track/Detect	28.47	25.61	32.83	22.87
2 Fixed Wing				
Visual ID/IFF	19.03	21.86	28.86	17.26
Handoff	27.75	28.11	28.69	22.18
Track/Detect	26.26	25.85	31.93	22.33

DATA ATTACHMENT C-2

FACTOR SCORES FOR ALL SUBJECTS LOS-F-H GENERIC			
	ID/IFF	Handoff	Track/Detect
One Rotary Wing			
Operator 1	-1.40	-0.60	0.50
2	-0.90	-1.80	-1.00
3	0.20	-1.30	-1.00
4	0.01	-1.30	-0.60
5	1.10	1.30	1.40
SME 1	0.09	-1.20	-1.80
2	--	--	--
3	--	--	--
4	0.80	0.50	0.70
5	0.90	-0.30	1.30
6	-1.10	-1.60	-0.80
7	1.30	1.00	1.30
8	0.10	-0.08	-0.30
9	0.90	-1.50	0.02
Two Rotary Wing			
Operator 1	-0.60	-0.50	0.80
2	-1.20	-1.40	-0.90
3	0.60	-1.20	-0.20
4	0.20	-0.02	0.20
5	1.70	1.60	1.80
SME 1	-0.30	-1.20	-1.70
2	--	--	--
3	--	--	--
4	0.90	0.30	0.80
5	1.40	1.20	1.80
6	-0.90	-1.30	-1.30
7	1.60	1.40	1.70
8	1.30	1.40	1.20
9	-0.10	-1.70	0.70
Two Fixed Wing			
Operator 1	-0.90	-0.10	0.70
2	-0.30	-0.90	-0.30
3	0.30	0.20	0.90
4	-0.20	-0.60	-0.09
5	1.70	1.90	2.00
SME 1	-0.09	-1.40	-1.60
2	--	--	--
3	--	--	--
4	0.70	0.60	-0.07
5	0.40	1.00	1.50
6	-1.30	-1.60	-1.40
7	0.90	1.00	1.60
8	1.20	1.20	1.40
9	0.20	-1.80	0.60

APPENDIX D

SUBJECTIVE WORKLOAD RATINGS OF THE LOS-F-H MOBILE AIR DEFENSE MISSILE SYSTEM IN A FIELD TEST ENVIRONMENT *

Susan G. Hill James C. Byers Allen L. Zaklad
Richard E. Christ

Abstract

The air defense system, the Line-of-Sight-Forward-heavy, or LOS-F-H, was involved in a field test in the summer of 1988 to examine selected concepts regarding tactics, doctrine, organization, and training. Four subjective workload assessment instruments were applied: Task Load Index (TLX), Subjective Workload Assessment Technique (SWAT), Overall Workload (OW), and the Modified Cooper-Harper (MCH). Individual assessments of mission segments were made by the three members of each of two crews and one replacement crew member. Jackknife factor analysis revealed the presence of only a single workload factor and indicated that the mean factor loadings formed a consistent ordering ($F(3,18) = 50.25$, $p < .0001$): TLX (.942), SWAT (.900), OW (.898), and MCH (.818). Analyses of variance also examined the effects of different variables on the workload factor scores; significant findings were discovered which reflected both on the system and the test. Regression analyses indicated a significant negative relationship between workload ratings and system performance. These findings as well as informal lessons learned are discussed in the context of the development and validation of a methodology for assessing workload.

INTRODUCTION

The air defense system, the Line of Sight-Forward-Heavy or LOS-F-H, has a primary requirement to engage low-altitude helicopters and fixed-wing threat aircraft, as part of the Forward Area Air Defense System. A Non-Developmental Item Candidate Evaluation (NDICE) was conducted in 1987 to select a "baseline" LOS-F-H from among four off-the-shelf candidates provided by various teams of contractors. The selected candidate was the system evaluated in the present study.

In the summer of 1988 a Force Development Test and Experimentation (FDTE) for this system was held at Fort Bliss, TX. The purpose of this field test was to examine tactics, doctrine, organization and training in relation to LOS-F-H. The test took place over a six-week period, from late May through mid-July, 1988, with the first five weeks comprised of four-hour missions and the last week of 48-hour missions. The present study, called the FDTE "Basic" study, looked at the applicability and usefulness of operator workload (OWL) ratings in the four-hour missions.

Purpose

The objectives of the present investigation were: (a) to explore the applicability of alternative OWL scales under the conditions characterizing field test evaluations, and (b) to evaluate operator workload during LOS-F-H operations.

METHOD

Subjects

The subjects were seven soldier-operators of the LOS-F-H. The operators included two radar operators (RO) who were also the mission commander/squad leader and five electro-optical operators (EO) who were "gunners". The EOs were lower ranking enlisted men (Private First Class and Specialists) and the ROs were non-commissioned officers with the rank of Sergeant. The operators

* This appendix contains a revised and condensed version of unpublished Technical Memorandum Number 5, prepared by the indicated authors in 1989. The sections of this appendix which address the relationship between workload ratings and system performance were taken from another unpublished manuscript: Byers, J. C. & Hill S. G. (1989), Comparison of subjective workload ratings to field test performance of the LOS-F-H mobile air defense system (Technical Memorandum Number 8).

were organized into two crews, with two EOs and one RO in one crew and three EOs and the other RO in the second crew. The ROs operated solely in that position, while the other crew members switched roles between EO and driver (DR).

All seven soldiers had participated previously in two related studies of workload (Bittner, Byers, Hill, Zaklad, & Christ, 1989, and Hill, Zaklad, Bittner, Byers, & Christ, 1988 – see Appendices C and B of this report, respectively). Hence, they were familiar with the concept, the OWL scales and the OWL data collectors.

Test Design

The FDTE was conducted using a test-fix-test design. This test design permitted a set of tactics, techniques, and procedures (TTP), defined as a battle drill, to be tested, then fixed based upon an analysis of the test data, then tested again. The TTP tested were step-by-step descriptions of what the crew must do to accomplish various mission segments.

Typically, Mondays were devoted to retraining TTP that had been changed from the previous week and testing some missile reload battle drills. On Tuesday through Thursday of each week, one crew was tested in the first of two daily 4-hour missions and the other in the second mission. These 4-hour missions consisted of the following series of mission segments: (a) prepare for road march (i.e., checking out the LOS-F-H system and processing the march order), (b) road march (i.e., move along an established roadway) to the selected site, (c) emplace the system at a predesignated battle site, and (d) conduct a one-hour acquisition and tracking (Acq/Track) battle drill (on four separate occasions, as a one-man operation). Fridays and the weekends were used to analyze the collected data and develop alternative TTP.

There were several operational variables of interest that were systematically changed over missions. These included: day and night missions, mission-oriented protective posture (MOPP) levels (which could vary both within and between successive missions), and countermeasures (including obscurants) which were used by threat aircraft during different passes. The intent was to systematically vary the combinations of factors presented to the crews. Upon occasion, however, the planned variation could not be implemented (e.g., the smoke generator was inoperable) and, therefore, did not take place.

The crews were rotated so they were used equally often in the first or the second of two scheduled daily missions. These were scheduled to start at 0800 in the morning and 1300 in the afternoon. The night missions were conducted similarly, but the engagements were scheduled to begin at 2000 for the early mission and 2400 for the late.

Procedure and Instruments

Prior to the first day of the FDTE, all subjects were briefed about the specific purpose of their participation in the workload assessment portion of the study and necessary procedures were completed for using the two multidimensional rating techniques.

The procedure for data collection was fairly constant throughout the FDTE Basic study. The OWL data collector would observe the Acq/Track engagement segment of a mission in real time via a four-camera, three screen video set up in an M109 van located at the mission site. Upon completion of a 1-hour Acq/Track mission segment or a reload exercise, the crew would return to the base camp area and proceed directly to a debrief trailer where OWL data were collected. During the first two weeks of the FDTE Basic study, workload ratings were made using each of the following four rating scales: (a) Task Load Index (TLX) (Hart & Staveland, 1987), (b) Subjective Workload Assessment Technique (SWAT) (Reid, Shingledecker, & Eggemeier, 1981), Overall Workload (OW) (Vidulich & Tsang, 1987), and (d) Modified Cooper-Harper (MCH) (Wierwille & Casali, 1983). During the final three weeks, ratings were made using only the TLX and OW techniques.

RESULTS

Analyses were conducted in five phases which respectively examined: (a) factor validity analysis of the workload measures; (b) workload in mission segments; (c) workload in the Acq/Track segment; (d) one-man operations; and (e) the relationship between workload ratings and system performance.

Factor Validity Analyses

Principal Component Analysis (PCA) was conducted using BMDP4M (Dixon, 1983) on 42 sets

of workload ratings obtained for all subjects and segments during the first two weeks. Each set included the global workload measures obtained from each of the four rating scales. (The mean and standard deviation of global workload ratings for each scale are in Data Attachment D-1 at the end of this appendix.) This analysis revealed a single component hereafter termed the **OWL factor,** which explained 79% of the total variance. The results of this initial analyses supported the view that the four workload scales essentially provide assessments of a single common factor. (The factor scores for each subject's workload judgments are in Data Attachment D-2.)

Jackknife PCAs were then conducted on the workload ratings data set in order to evaluate the stability of the factor loadings of the four scales (i.e., correlations with the OWL factor). Jackknife analysis generally involves successive analyses (PCAs in the present case) dropping subjects one-at-a-time from a data set in order to examine the stability of parameter estimates (Hinkley, 1983). In the present case, with four factor loadings and the 7 subjects, a 4 (loadings) by 7 (subjects dropped) matrix was produced which could be analyzed by conventional repeated measures analysis of variance (ANOVA). The ANOVA (using BMDP2V in Dixon, 1983) revealed a significant difference between the workload scale factor loadings ($F(13,18) = 50.25$, $p < 0.0001$). Subsequent analysis revealed the following ordering of the mean factor loadings:

TLX(.942), SWAT(.900), OW(.898), MCH(0.818).

All differences are significant, with the exception of SWAT-OW.

For the remaining four weeks of testing, only TLX and OW ratings were obtained. The **OWL factor scores** which were the basis for the workload analyses in the following sections were derived from a PCA of the TLX and OW scores collected during the five weeks of testing of four-hour missions.

Workload in Mission Segments

The amount of workload experienced by different LOS-F-H crew members during different mission segments was investigated by ANOVA. The **OWL factor scores** were used as the workload score. The segments examined are described as: Acquisition/Tracking (Acq/Track), Emplacement, Reload, One-man Operations, and Road march.

A crew member position main effect was found ($F(2,238) = 55.19$, $p < 0.00018$). As may be seen in Table D-1, the DR has the least workload (-1.04), while EO (0.18) and RO (0.49) had greater workload. The differences between EO and RO were insignificant, while the differences between DR and EO, and DR and RO were significant. The mission segments were found to be significantly different ($F(4,199) = 9.38$, $p < 0.0001$). As may be seen in Table D-1, the greatest workload is reported for One-man Acq/Track Operations and the least for Road March.

The joint effect of crew position and mission segments on workload was separately analyzed for the three segments of Acq/Track, Emplace, and Reload. These three segments were rated by subjects in all three crew positions (one-man Acq\Track operations and driving the

Table D-1

OWL Factor Scores for Mission Segments and Crew Member Positions

	RO			EO			DR			ALL POSITIONS		
MISSION SEGMENT	Mean	SD	n	Mean	SD	n	Mean	SD	n	Mean	SD	n
Acq/Track	0.39	0.63	58	0.18	0.88	61	-1.13	0.41	29	0.01	0.91	148
Emplace	0.61	0.33	11	-0.82	0.67	11	-1.05	0.47	11	-0.42	0.90	33
Reload	1.13	0.65	5	1.74	1.18	5	-0.31	1.54	3	1.03	1.29	13
One-man Ops.	1.14	0.69	2	1.45	1.75	2	----	----	--	1.30	1.10	4
Roadmarch	----	----	--	----	----	--	-0.99	0.18	6	-0.99	0.18	6
ALL SEGMENTS	0.49	0.62	76	0.18	1.05	79	-1.04	0.54	49	0.00	1.00	20[illegible]

vehicle in road march each were rated by one only one subject per mission). A significant Position x Segment interaction was found ($F(4, 185) = 5.42$, $p < 0.0004$). This can be seen in Table D-1. The DR indicates less than average workload in all three segments. Both the RO and EO report higher than average workload for the Acq/Track and Reload segments. However, the RO has higher than average workload while the EO has much lower than average workload during emplacement.

The TLX subscales ratings for position-by-mission segment are presented in Figure D-1. The height of the stacked column represents the total workload for the three segments of Acq/Track, Emplace, and Reload. Examination of the figure shows the differences in types of workload experienced in various mission segments by position. For example, in Acq/Track, the RO experiences more total workload than the EO (although not significantly different), although the EO experiences more temporal demand than the RO. Another example is that there is substantially larger Physical and Temporal Demand components and a larger Effort component (showing how hard someone is working) for the Reload than any other mission segment. Figure D-1 also shows that the RO always has larger Performance subscale scores (i.e., he perceives he has been less successful in accomplishing his task) than either the EO or DR.

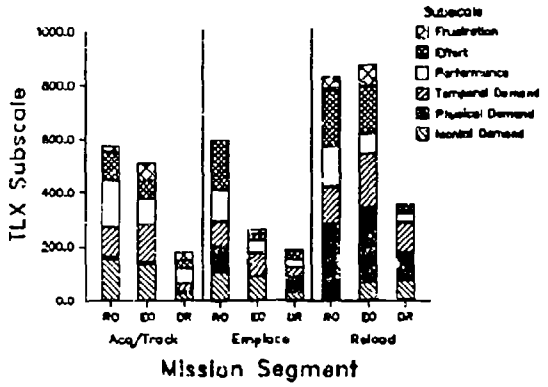


Figure D-1. *The effect of mission segment and crew member position on TLX subscale ratings.*

Workload Within the Acq/Track Mission Segment

Effects of specific tasks. Workload given by OWL factor scores was examined for specific tasks in the Acq/Track mission segment. The combination of a specific task and the crew member who performs the task is called an event. The Acq/Track events which were rated in this study include: (a) for all three crew members, the Entire Acq/Track Mission Segment; (b) for the RO, the four events defined by Detecting and Acquiring both Fixed- and Rotary-Wing aircraft; and (c) for the EO, the four events defined by Acquiring and Tracking of both Fixed- and Rotary-Wing aircraft. There was no significant difference among these workload ratings, due, in part, to large variations in the ratings over subjects and missions. However, there were two potentially meaningful trends evident in these data. First, the workload reported by an RO performing his specific Acq/Track tasks was generally higher than those reported by an EO doing his tasks (0.39 and 0.18, respectively). Second, workload scores of the EO for Acquiring and Tracking Fixed-Wing aircraft (0.04 and 0.28, respectively) were higher than for Acquiring and Tracking Rotary-Wing aircraft (-0.23 and -0.27, respectively).

Effects of mission variables. The effect of various mission variables on Acq/Track event workload was examined. Although the mean OWL factor scores for variation in MOPP Level suggest that more workload was experienced in MOPP 4 (0.16) than in MOPP 0 (-0.05), the difference was not significant. Similarly, no significant differences were found between clear viewing conditions (-0.04) and those obscured by smoke (0.15), or between conditions in which the crew was or was not alerted by outside elements that a target was entering its sector (-0.12 and 0.13, respectively). A difference was found in rated workload between day and night missions ($F(1,146) = 3.50$, $p < 0.06$). Day missions were rated as having more workload (0.10) than night missions (-0.21), perhaps due to the elevated temperature during day-time missions in the desert test environment.

Workload During One-man Acq/Track Operations

One-man Acq/Track operations were performed during four missions of the FDTE. Two ROs and two EOs participated in these missions. A separate ANOVA of these missions revealed no significant effects due to crew member duty position, Acq/Track event, or TLX subscale. There was a tendency, however, for ROs to report higher levels of global workload with the TLX for these operations than EOs (46.2 and 30.3, respectively). The largest difference between the RO and EO is for the task of "Tracking Fixed-Wing," for which the EOs are practiced and the ROs are not. The only

event that ROs rated as having less workload than the EOs was "Detecting Fixed-Wing," for which the RO was much more practiced (using the radar scope) than the EO.

The Relationship Between Workload Ratings of Individual Crew Member and System Performance

The OWL factor scores derived for each crew member when they rated specific tasks or events in each one-hour Acq/Track mission segment included one defined as "Entire Acq/Track Mission Segment." These specific scores were compared to a measure of system performance for the corresponding missions. The system performance data were provided by the U.S. Army Air Defense Artillery Board at Fort Bliss, Texas. This agency was responsible for the conduct of the LOS-F-H FDTE.

The baseline system performance measure (PERFORM) used the percentage of successful engagements during aircraft passes over the entire FDTE basic study. This percentage was obtained by dividing the number of passes scored "successful" by the test agency by the total number of passes scored. (Passes counted as "No Test," for any reason, were not included.) Other performance measures were derived from the baseline data. These measures were formed by withholding certain types of passes from the total number scored. For example, since workload ratings are associated with an operator's experiences, his perceived workload would not be affected if he was unaware of the existence of an aircraft. Therefore, one such alternative measure eliminated from consideration all passes scored as "did not detect target." Analyses with these alternative system performance scores did not reveal any meaningful relationships that were not also found with the baseline PERFORM data.

A stepwise regression with PERFORM as the dependent measure and independent measures of the RO OWL factor scores (based on TLX and OW ratings only) and dichotomous (dummy) variables to index the two ROs making the ratings stopped after the accretion of only the OWL factor score variable. This analysis revealed a significant correlation, $R = -0.65$ ($F(1,48) = 34.5$, $p < 0.001$). Similar analyses for EO, DR, and all positions combined revealed no significant relationship between PERFORM and OWL factor scores. A graphical representaion of the significant regression of PERFORM onto OWL factor scores of the ROs is given in Figure D-2.

Stepwise regressions with PERFORM as the dependent variable and the TLX Performance subscale ratings as the independent variable

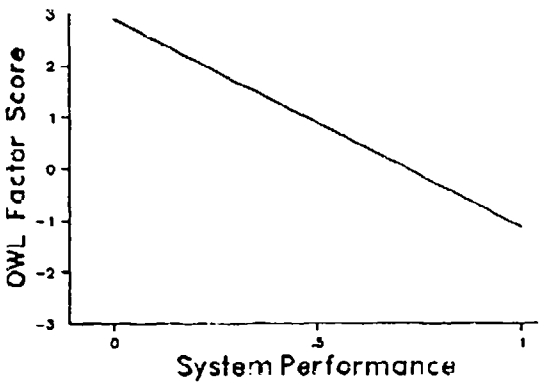


Figure D-2. *The relationship between workload ratings of ROs and system performance.*

revealed significant relationships. (The workload rating on the Performance subscale is given its highest value when a subject perceives that his or her performance was a complete failure and its lowest value when performance is judged to be perfect). The result for the RO position was similar to the one reported above, $R = 0.56$, ($F(1,67) = 31.03$, $p < 0.001$. Similar analyses using TLX subscale ratings from crew members in the EO position revealed a significant multiple correlation, $R = 0.65$, $F(3,66) = 16.14$, $p < 0.001$. There was no significant relationship between system performance TLX performance subscale ratings provided by the DR.

DISCUSSION

Factor Validity

An ordering of the factor validities of the four measures resulted in TLX > SWAT > OW > MCH. The ordering is somewhat familiar to those found in earlier studies (e.g., Bittner et al., 1989, and Hill et al., 1988 -- see also Appendices C and B, respectively). These results support previous conclusions that TLX had the highest factor validity.

Workload in Mission Segments

Workload was examined as a function of

mission segments. Clearly, the DR has very little workload, while the RO and EO had about the same workload across all segments, save Emplacement. The RO and EO workload scores were highest for the Reload and One-man operation mission segments (see Table D-1). The subscale analysis (Figure D-1) was particularly interesting, suggesting the different dimensions which contributed to OWL for the different positions. The Acq/Track mission segment had the greatest mental demand while Reload had the strongest physical, temporal, and effort components. The emplacement mission shows a large position-by-subscale interaction (Figure D-1), with the RO experiencing the greatest overall OWL, although his mental and temporal demand are similar to those reported by the EO. These effects of mission segment and duty position correspond well with expectations and observation, suggesting substantial face validity of the composite and subscale ratings.

Workload During Acq/Track Segments

The results indicate no significant differences in workload across position (RO and EO) and task during Acq/Track segments. Of the mission variables, only day/night had a significant effect on workload. This is somewhat surprising. In particular, it was thought that MOPP level would affect workload. However, there was no difference. The workload ratings may reflect a lower level of work being done because of the heat.

One-Man Acq/Track Operations

It is difficult to make any firm conclusions based on only four one-man missions. Indeed there are only two missions for each of the two duty positions. However, the One-man Operations segment has the highest average OWL score (1.30). The RO has greater OWL scores than does the EO, perhaps because the RO feels more responsible and the EO knows he is not expected to do well so he feels relatively relaxed.

The Relationship Between Workload Ratings and System Performance

The significant correlations found between operator workload ratings and system performance were in accordance with expectations. That is, the results indicate decreasing system performance with increases in operator workload (OWL factor score or TLX Performance subscale score). The strongest correlations were found when analyzing data for the RO position. Possible reasons for this include: (a) the ROs had the highest average workload rating for the Acq/Track mission segment and may have been more susceptible to performance decrements when workload increased; (b) the ROs, with both radar knowledge and a view of the EO's display, may have the most accurate opinion of how the system and crew is performing, which may influence TLX performance subscale ratings; and (c) greater experience and age may make the ROs more perceptive raters of workload.

The results for the EO and Driver positions are more problemmatic. Considering the Driver's role during an engagement mission (i.e., with very little to do, the Driver sometimes slept) and the low workload ratings by those in the Driver position, the expectation was that changes in Driver workload would have no effect on system performance. The expectation for the EO position, given the imporatant role that the EO has in the engagement sequence, was that operator workload would correlate with system performance. The TLX Performance Subscale analysis agreed with expectation while the OWL factor score analysis did not.

CONCLUSIONS

Subjective ratings of operator workload in the LOS-F-H FDTE indicated:

(1) Global workload ratings were much greater for the RO and EO than for DR,

(2) Some significant effects of mission variables on workload,

(3) Differences in both magnitude and dimensions of workload among mission segments, and

(4) Increases in operator workload are associated with decreases in system performance.

Analyses revealed meaningful results with substantial face validity.

REFERENCES

Bittner, A. C., Jr., Byers, J. C., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1989). Generic workload ratings of a mobile air defense system (LOS-F-H). Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1476-1480). Santa Monica, CA: Human Factors Society.

Byers, J. C., & Hill, S.G. (1989). Comparison of subjective workload ratings to field test performance of the LOS-F-H mobile air defense system (Technical Memo 8). Willow Grove, PA: Analytics, Inc.

Dixon, W.J. (Ed.). (1983). BMDP statistical software. Los Angeles, CA: University of California Press.

Hart, S.G. & Staveland, L.E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.S. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Hill, S. G., Zaklad, A. L., Bittner, A. C., Jr., Byers, J. C., & Christ, R. E. (1988). Workload assessment of a mobile air defense missile system. Proceedings of the Human Factors Society 32th Annual Meeting (pp. 1068-1072). Santa Monica, CA: Human Factors Society.

Hinkley, D.V. (1983). Jackknife methods. In S. Kotz, N.L. Johnson, & C.B. Read (Eds.), Encyclopedia of statistical sciences: Vol.4 (pp. 280-287). New York: Wiley.

Reid, G.B., Shingledecker, C.A., & Eggemeier, T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting (pp. 522-525). Santa Monica, CA: Human Factors Society.

Vidulich, M.A. & Tsang, P.S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

Wierwille, W.W. & Casali, J.G. (1983). A validated rating scale for global mental workload measurement application. Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.

DATA ATTACHMENT D-1

COMPARISON OF WORKLOAD RATING SCALES FOR LOS-F-H BASIC STUDY

MISSION SEGMENT/ POSITION	RATING SCALE			
	OW	MCH	SWAT	TLX
	---------- MEANS ----------			
MISSION SEGMENT				
Acq/Track	35.00	23.52	32.52	31.61
Emplace	28.00	--	--	23.47
Road March	21.42	--	--	13.83
Reload	54.61	18.33	20.86	49.38
One-Man Ops	57.50	--	--	37.45
POSITION				
RO	43.81	19.80	39.13	41.41
EO	38.53	40.73	49.04	32.15
DR	16.27	6.76	0.83	13.17
	---- STANDARD DEVIATIONS ----			
MISSION SEGMENT				
Acq/Track	19.22	23.15	32.02	15.96
Emplace	18.63	--	--	14.62
Road March	13.13	--	--	5.48
Reload	25.20	12.70	16.99	23.61
One-Man Ops	18.93	--	--	23.06
POSITION				
RO	13.13	11.90	26.19	13.78
EO	22.45	26.95	31.06	18.83
DR	12.68	8.44	0.77	9.17

DATA ATTACHMENT D-2

		CREW MEMBER POSITION RO	EO	DR
	RELOAD MISSION			
100		0.78	1.56	-1.00
101		2.02	3.09	1.47
102		1.58	2.62	-1.00
103		0.77	1.34	--
104		0.46	0.08	--
	BASIC MISSION/EVENT			
321	Entire Mission	1.22	-1.00	1.34
331	Entire Mission	0.70	2.55	-0.61
332	Entire Mission	1.16	-1.00	1.31
421	Entire Mission	-0.09	1.39	-0.69
	Detect FW	0.04	--	--
	Track FW	--	1.91	--
	MSCS	--	--	-0.19
422	Entire Mission	0.97	0.36	-1.00
	Detect FW	0.92	--	--
	Track FW	--	-1.00	--
	MSCS	--	--	-0.78
432	Entire Mission	1.07	-0.42	-1.00
	Detect FW	0.36	--	--
	Track FW	--	0.40	--
	MSCS	--	--	-1.00
441	Entire Mission	-0.10	1.39	-1.00
	Detect FW	0.26	--	--
	Track FW	--	1.32	--
442	Entire Mission	0.97	-0.62	-1.00
	Detect FW	0.79	--	--
	Track FW	--	0.56	--
511	Entire Mission	0.55	0.02	-2.00
	Detect RW	1.06	--	--
	Acquire RW	0.82	-0.80	--
	Track RW	--	-1.00	--
531	Entire Mission	0.80	-0.16	-0.73
	Detect RW	1.03	--	--
	Acquire RW	1.36	-0.36	--
	Track RW	--	-0.38	--
	Listening for MSCS	--	--	-1.00
	Plotting MSCS	--	--	-0.76
	Emplacement	0.66	-0.62	-1.00

DATA ATTACHMENT D-2 (Continued)

		CREW MEMBER POSITION RO	EO	DR
	BASIC MISSION/EVENT			
532	Entire Mission	1.53	1.56	-1.00
	Detect RW	0.71	--	--
	Acquire RW	0.42	-0.55	--
	Track RW	--	-1.00	--
	Listening for MSCS	--	--	-1.00
	Plotting MSCS	--	--	-1.00
541	Entire Mission	0.88	0.27	-0.78
	Detect FW	1.17	--	--
	Acquire FW	1.10	-0.18	--
	Prioritize Targets	0.38	--	--
	Track FW	--	-2.00	--
	Emplacement	0.90	-2.00	-0.83
	Driving	--	--	-0.82
542	Entire Mission	-0.09	1.34	-0.95
	Detect FW	-0.38	--	--
	Acquire FW	0.11	1.07	--
	Prioritize Targets	-0.23	--	--
	Track FW	--	0.73	--
	Choose Target Mode	--	-0.39	--
	Emplacement	0.22	-0.37	-0.39
	Driving	--	--	-0.81
621	Entire Mission	0.81	1.38	-0.46
	Detect FW	0.05	--	--
	Track FW	--	1.10	--
	Detect RW	0.41	--	--
	Acquire RW	0.95	0.14	--
	Track RW	--	0.86	--
	Acquire FW	0.25	0.55	--
	Emplacement	1.04	-0.16	-0.56
622	Entire Mission	0.54	-0.99	-2.00
	Detect FW	0.97	--	--
	Track FW	--	-1.00	--
	Acquire FW	0.63	-0.35	--
	Emplacement	0.53	-1.00	-2.00

DATA ATTACHMENT D-2 (Continued)

	BASIC MISSION/EVENT	CREW MEMBER POSITION RO	EO	DR
632	Entire Mission	0.65	-0.29	-2.00
	Prioritize Targets	0.62	--	--
	Choose Target Mode	--	-0.03	--
	Hangfire	-0.23	-0.84	--
	Emplacement	--	-0.91	-1.00
	Driving	--	--	-1.00
721	Entire Mission	0.61	-0.27	-2.00
	Detect FW	0.30	--	--
	Track FW	--	-0.82	--
	Detect RW	0.95	--	--
	Acquire RW	0.40	-0.54	--
	Track RW	--	-0.91	--
	Acquire FW	0.29	mis	--
	Listening for MSCS	--	--	-2.00
	Plotting MSCS	--	--	-2.00
	Choose Target Mode	--	-0.74	--
	EO Target Detect/Engage	--	-0.66	--
	Emplacement	0.26	-0.95	-2.00
722	Entire Mission	1.22	0.98	-0.85
	Detect FW	-0.31	--	--
	Track FW	--	0.06	--
	Detect RW	-0.65	--	--
	Acquire RW	-0.50	0.32	--
	Track RW	--	-0.46	--
	Acquire FW	-0.56	0.34	--
	Choose Target Mode	--	0.83	--
	Hangfire	-0.60	-0.01	--
	EO Target Detect/Engage	--	0.68	--
	Emplacement	1.24	-0.17	-1.00
741	Entire Mission	-0.01	0.94	-1.00
	Detect FW	-0.76	--	--
	Detect RW	-0.90	--	--
	Prioritize Targets	-0.53	--	--
	Trouble Shooting	-1.00	--	--
	Driving	--	--	-1.00
	Track FW	--	0.26	--
	Acquire RW	--	0.19	--
	Track RW	--	0.46	--
	Acquire FW	--	0.51	--
	Target Recognition	--	0.67	--

APPENDIX E

SUBJECTIVE WORKLOAD ASSESSMENT DURING 48 CONTINUOUS HOURS OF LOS-F-H OPERATIONS

Susan G. Hill James C. Byers Allen L. Zaklad
Richard E. Christ

Abstract

Two operator workload (OWL) rating scales were used to obtain judgments of OWL throughout 48 continuous hours of operation of the LOS-F-H air defense system. The Task Load Index (TLX) and Overall Workload (OW) scales were administered to two crews in two different 48-hour operations. Results indicate that workload increases significantly over time. Regression analyses suggest that OWL scores can be described as a combination of hour into the mission and job being performed. These findings are discussed in the context of the development and validation of a methodology for assessing OWL.

INTRODUCTION

The air defense system, the Line of Sight-Forward-Heavy or LOS-F-H, has a primary requirement to engage low-altitude helicopters and fixed-wing threat aircraft as part of the Forward Area Air Defense System. A Non-Developmental Item Candidate Evaluation (NDICE) was held in Fall, 1987, and the winning system was chosen as the Army prototype LOS-F-H. Initial OWL assessments of the winning candidate were conducted retrospectively, by asking the soldier-operators to make judgments of OWL by viewing videotapes of their own performance during NDICE (Hill, Zaklad, Bittner, Byers, & Christ, 1988) and to make overall judgments of various generic mission segments and tasks (Bittner,Byers, Hill, Zaklad, & Christ, 1989) -- see Appendices B and C of this report, respectively.

A Force Development Test and Experimentation (FDTE) program for the LOS-F-H system was held in June-July, 1988 at Fort Bliss, TX. During this FDTE, OWL assessments of various tasks under a varied of mission contexts were obtained using a family of subjective OWL ratings (Hill, Byers, Zaklad, & Christ, 1989 -- see Appendix D of this report). Following five weeks of two four-hour missions per day, the FDTE examined performance in a 48-hour mission designed to emulate the operational mode summary for the LOS-F-H. This paper describes the 48-hour operations, the methodology and procedures used to obtain OWL assessments, and the results and discussion of the OWL assessment.

Purpose

The objectives of the present investigation were: (a) to explore the applicability of the OWL scales for obtaining workload assessments during 48-hour continuous operations; (b) to evaluate the relationship between mission variables and the workload assessments of the crew members; and (3) to compare the results of the present programmatic investigation with those from earlier efforts in the series.

METHOD

Subjects

Two three-member crews participated, one crew in each of the two 48-hour missions. The three crew positions are radar operator (RO), electro-optical operator (EO) or "gunner" and a driver (DR). Each crew member had some cross-training for all positions; however, the RO remained the same person throughout the 48 hours

---

* *This appendix contains a revised and condensed version of a paper presented at and published in the Proceedings of (pp 1129-1133) the 33nd Annual Meeting of the Human Factors Society.*

(with one exception in one crew) while the EO and DR switched positions after the first 24 hours (with one exception in one crew). The two exceptions occurred when: (a) the RO did not participate in a mission and (b) the scheduled EO was temporarily removed from the test and the scheduled DR participated as EO.

The EO/DRs were junior level enlisted men and the ROs were Non-Commissioned Officers (NCOs). These same crews participated in previous field tests of the LOS-F-H system; they had just completed five weeks of testing for four-hour missions. Consequently, the operators were experienced with the OWL scales and with being observed. They were also sensitive to OWL concerns and comfortable with the data collectors.

Procedure and Instruments

At periodic times during the 48 hours, the crew was asked to give OWL ratings. Two rating scales were used to obtain OWL rating: Task Load Index (TLX), Hart & Staveland, 1987, and Overall Workload (OW), Vidulich & Tsang, 1987. At one data collection interval, only the TLX scale was used. Based on the results from several previous studies in this series it was decided that global workload measures would be obtained with the TLX scale by computing the arithmetic mean of the ratings given to the six subscales to generate a "raw" TLX score (RTLX), rather than the weighted average of the subscale ratings. It has been shown by Byers, Bittner, and Hill (1989) that the two approaches to computing a global score from the subscale ratings yielded essentially identical results. A desirable consequence of using the RTLX is that no paired-comparison weights need to be obtained for each task whose workload was being evaluated.

During the rest of the mission, the data collector made notes as to crew activities and attitudes to the degree that the crew could be observed. An OWL data collector was on site at all times, with the exception of 0000 to 0530, when the system was off and the crew slept. Two formal debriefs of the crew took place. The first took place in the field after the first 24 hours during an administrative break in the mission. The second debrief took place in a debriefing trailer at the base camp after the completion of the 48-hour mission.

The two different 48-hour missions were conducted at different times. However, the schedule of events planned for both missions were the same, and included 14 Road March, eight Acquisition/Tracking (Acq/Track), and six Missile Reload mission segments. With only the exception of two canceled reload segments, all events took place approximately as scheduled. Each of the two missions were scheduled to began at 1200 on the first day and continue to 1200 of the third day; the system was shut down from 0000 to 0530 on the second and third day, during which time the crews were scheduled to sleep. In terms of physical conditions, the days were very hot and the evenings were cool. The crew compartment of the weapon system had no air conditioning and there was great concern about heat stress on the crew, particularly during the day and when in full chemical protective posture.

The OWL measures consisted of a rating of the workload of the "Overall Mission so far," or a cumulative assessment of workload. It was decided that a cumulative assessment was better than a judgment of workload since the last rating because the ratings might be hours apart and thus lessen accuracy. At the 24 and 48 hour debriefs, additional OWL ratings were obtained on engagement-specific tasks. At the conclusion of the 48 hours, OWL ratings were obtained from the two junior ranking crew members on "Your 24 hours as EO" and from all three crew members on the "Entire 48-hour mission."

RESULTS

Quantitative analyses were conducted in three phases which respectively examined: (a) the relationship between the two workload scales, (b) the effect of time on workload, (c) the relationship of workload to mission variables. The analyses examined the two crews separately as well as both crews together. In many cases, the two different sets of crew members experienced variations in the exact timing of scheduled events and in environmental conditions. Consequently, it was decided that combining them would be less useful than examining them separately. Descriptions of the data obtained during two debriefs of the crews (held at 24 and 48 hours into the mission) are reported separately in the qualitative analyses section.

Quantitative Analyses

Factor analysis. Principal components analysis (PCA) on OW and raw (unweighted) TLX (RTLX) ratings was performed using the BMDP4M statistical software package (Dixon, 1983). A single factor, hereafter called the **OWL factor**, was found which explained 82% of the total variance. These results support the view that the two workload scales essentially provide assessments of a single common factor. The resulting **OWL factor scores** were used in the workload analyses reported in the following sections. (The OWL factor scores for each subject's workload ratings are in Data Attachment E-1 at the end of this appendix.)

Effects of time on workload. The workload ratings were divided into different time blocks to examine the effect of time on workload. An attempt was made to make divisions such that each block contained events that potentially would affect workload. The two crews were examined separately because of the differences between missions (as mentioned previously) and because there were a different number of workload measurements made over the 48 hour period. There are more opportunities to obtain ratings from the second crew than from the first crew.

The workload scores were first examined by day. For both Crews 1 and 2, Day 1 workload ratings were significantly different from Day 3 ratings, with workload higher at the end of the mission ($F(2,18) = 5.07$, $p < 0.018$; $F(2,27) = 12.42$, $p < 0.0002$). The means ratings for Crew 1 are -0.72 and 0.66 for Days 1 and 3, respectively. Corresponding mean ratings for Crew 2 are -1.31 and 0.33. When the mission was examined in greater detail (i.e., seven time blocks for Crew 1 and nine time blocks for Crew 2), there was a significant effect of time for both crews ($F(6,12) = 6.00$, $p < 0.0042$; $F(9,18) = 3.11$, $p < 0.02$. The mean rating for each time block for each crew are shown in Figure E-1. These workload scores are graphically illustrated or plotted as a function of hour into the mission for each crew. As can be seen, the crews report the same general increase in workload across time (with the primary exception of a decreased OWL score for Crew 2 at Hour 7 into the mission).

Position effects. Crew member position significantly affected the OWL factor scores. In particular, the RO had a greater average workload that either EO or DR (RO = 0.20; EO = 0.18; and, DR = -0.7 respectively). Also, results suggest that there is more workload involved with being the EO during the second 24 hours that being EO during the first 24 hours ($F(1,2) = 26.9$, $p < 0.0035$). The means are -0.21 for the EO in the first 24 hours and 1.45 for EO in the second 24 hours.

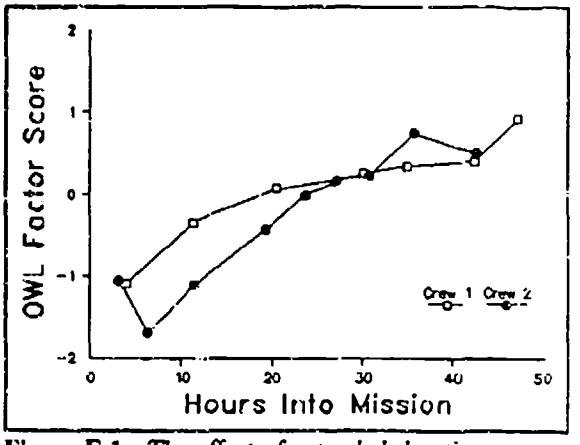


Figure E-1. *The effect of extended duration missions on workload.*

Effects of mission variables. Regression analyses were used to examine the relationship of workload to various mission variables. The variables of interest were: time of day (i.e., day or night), time from last sleep period, time from last reload, time from last Acq/Track segment, time since last MOPP 4 condition, time into mission (to the nearest quarter hour), and job (i.e., whether they were performing an active job (RO or EO) or an inactive job (DR)). Regression was performed for each crew separately and for both sets of crew data together. The resulting regression equations for Crew 1 only and both crews together were quite similar, workload being related to the same two factors of hours into the mission and job ($R = 0.83$ and 0.81, $n = 21$ and 48, respectively). The equation obtained for the data of both crews together is:

OWL = -1.964 + (0.049 * Hour) + (0.928 * Job).

Two additional factors entered the regression equation for the data of Crew 2 only: time since last MOPP 4 condition and a measure of physical symptoms. Crew 2 did have the occurrence of two heat-related incidents which did not occur for Crew 1. It may be that the additional factors entering the Crew 2 equation were due to these heat-related incidents. If so, and if the occurrence of such incidents are rare, the regression equation shown above may be the best description of the

relationship between mission variables and workload.

Qualitative Analyses

Two debriefs (at 24 and 48 hours) provided direct, qualitative information from the operators. Although interview data are difficult to analyze, they are reported here in an effort to provide a basis for interpreting the reported quantitative results. Few specific comments directly regarding workload were made.

After 24 hours. During the first 24 hours, the two ROs (i.e., squad leaders) got 1 - 2 hours of sleep. The EO/DRs received 2.5 - 3.0 hours sleep. General comments indicated that the first 24 hours were pretty much as expected and that the next 24 hours would be about the same. The crews reported that in some ways they felt more relaxed in this extended operational scenario and not as rushed to accomplish preliminary placements and setups as they had been during the four-hour missions that had been experienced in the preceding 5 weeks of this field test. The crews also indicated that they felt that the Acq/Track missions during the first 24 hours of this extended mission were not as difficult as those experienced during the shorter operations. Some complaints were made regarding MOPP 4 gear (hard to see out of mask; very draining); missile reloads (flying insects bothered the crew during night operations); and other matters (e.g., cramped quarters inside the fire unit).

Some potentially important comments were made regarding crew organization. As mentioned previously, the operator assigned as EO remained in that position for the first 24 hours. In both crews, the "first" EO remarked that it was very difficult to remain as EO for the first 24 hours (which included 4 Acq/Track missions) because the electro-optics display screen is difficult to look at continuously. These EOs claimed the extended requirement for viewing the display screen caused eyestrain and headaches. The operators suggested switching positions more often. The drivers concurred with this suggestion because they felt their job was very boring over a 24 hour period.

During one Road March, Emplacement, and Acq/Track mission, the squad leader also drove the vehicle while the other two crew members acted as EO and RO. The reason for the position change was to try out a new organizational concept (see Hill, Byers, Zaklad, Bittner, & Christ, 1989 or Appendix F of this report for details). At the debrief, the RO/squad leader stated that he felt demoted by having to drive (traditionally, the driver is the lowest ranking member of the crew), but liked the ability to see outside of the vehicle which can only be done from the driver's position.

Although other comments were made during the debrief, those presented above give the primary areas discussed and the opinions of the crews.

After 48 hours. The soldiers reported the total sleep they received during the 48-hour period as 8 and 13 hours for the squad leaders, and 8, 10, 10 and 13 for the other crew members. One crew had the 5-gallon water container refilled three times while the other crew had the container refilled four times. Although it is not known precisely how much water was consumed, it can be inferred that each crew member had approximately 5 gallons over the 48-hour period.

General comments made at the 48-hour debrief included that the experience was easier than expected and the soldiers felt more relaxed after they had been on the system for a longer period of time (one expressed it as feeling "at home"). The crews felt that wearing MOPP 4 gear was their most difficult experience during the 48 hours because of the heat; the system is just too hot inside to wear MOPP 4.

Crews reported that vibration, noise or riding sideways in the vehicle were not problems. They felt that Identification Friend or Foe (IFF) and early warning from Manual Shorad Control System (MSCS) both enhanced the operators abilities to successfully engage targets.

Several comments were made regarding missile reload operations. The crews felt reloads were demanding and draining physically and too many had been scheduled for the 48 hours. Reloads at night presented some unusual problems. For example, one RO felt as if he might fall off the top of the vehicle because he couldn't see very well.

Again, those operators who had served as EOs for 24 hours reiterated the demanding nature of watching the electro-optics display screen for several missions and their desire to switch positions more often than they had during this 48 hour period.

DISCUSSION

Several issues need discussion. First is the basic question of sample size. All the analyses presented are based on two crews of three members each. This is not a large sample from which to draw strong conclusions. However, it is believed that these were representative crews and the results certainly present a reasonable picture of operator workload during these 48-hour missions.

Workload

An important unresolved issue is what exactly to measure when investigating workload across time. The measure used here was to ask for workload ratings of the "Mission So Far." Perhaps some other measure would have been more appropriate. Similarly, ratings were obtained after a significant event had occurred and when circumstances permitted. Would it be more appropriate to obtain measures at fixed intervals (e.g., every three hours) regardless of event occurrence? These issues deserve some thought and attention.

Another issue is how to interpret the OWL ratings obtained and analyzed. If the label "Mission So Far" is taken literally, then the scores should be cumulative across time and always be increasing. Even if no workload was experienced since the previous measurement, the cumulative workload would, at least, stay the same. However, although the trend was increasing for both crews, there were a couple of points where the workload "so far" decreased. Another interpretation would be that at each measurement, an averaging of the workload for the "mission so far" is taking place. This fits the results somewhat better. For example, if there is about the same or increasing workload, an average will increase across time. However, if the workload in the latest period is particularly low, an average across time will show a decrease in the reported workload.

There is also the possibility that there were beginning or end of mission effects. For example, the crews may have been apprehensive about participating in 48 hour operations and initially rated workload high. As a little time passed, and things were not as bad as the crew had thought they might be, the workload rating was lessened. This might explain the OWL score for Crew 2 at Hour 7. Similarly, as the end of the mission approached, crews may have differentially perceived workload influenced by the end of the mission itself.

The workload results obtained from this study support previous conclusions that the RO and the EO have much greater workload than the driver (cf., Hill, Byers, Zaklad, & Christ, 1989).

Effects of Mission Variable

The significant factors used to predict workload were the hour into the mission and the job being performed. The importance of the hour is not surprising, the OWL score appears to be an average across time and would tend to increase the longer the mission lasts. Workload may also be associated with fatigue. The importance of job in the regression equations suggests the large difference in workload between the positions as discussed previously. The additional factors of MOPP and physical symptoms in the Crew 2 regression equation are believed to be associated with the particular heat incidents that took place. These relationships are interesting, but a larger sample should be collected and analyzed before any firm conclusions are made.

CONCLUSION

Based on the limited sample available, workload ratings were affected across time. Although questions remain concerning the most appropriate way to measure workload over extended periods, the results and suggested interpretations presented here are promising and future workload investigations during extended missions should be pursued.

REFERENCES

Bittner, A. C., Jr., Byers, J. C., Hill, S. G., Zaklad, A. L. and Christ, R. E. (1989). Generic workload ratings of a mobile air defense system (LOS-F-H). Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1476-1480). Santa Monica, CA: Human Factors Society.

Byers, J. C., Bittner, A. C., Jr. and Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? Advances in Industrial Ergonomics and Safety, Vol. 1 (pp. 481-485). London: Taylor and Francis.

Dixon, W. J. (Ed.). (1983). BMDP statistical software. Los Angeles, CA: University of California Press.

Hart, S. G. & Staveland, L. E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. S. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Hill, S. G., Byers, J. C., Zaklad, A. L., Bittner, A. C., Jr., & Christ, R. E. (1989). Prospective workload ratings of LOS-F-H mobile air defense missile system (Technical Memo 2). Willow Grove, PA: Analytics, Inc.

Hill, S. G., Byers, J. C., Zaklad, A. L., & Christ, R. E. (1989). Subjective workload ratings of the LOS-F-H mobile air defense missile system in a field test environment (Technical Memo 5). Willow Grove, PA: Analytics, Inc.

Hill, S.G., Zaklad, A.L., Bittner, A.C., Jr., Byers, J. C., & Christ, R.E. (1988). Workload assessment of a mobile air defense missile system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1068-1072). Santa Monica, CA: Human Factors Society.

Vidulich, M.A. & Tsang, P.S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

DATA ATTACHMENT E-1

FACTOR SCORES FOR THE LOS-F-H 48-HOUR STUDY

MISSION 1

DATE/TIME	RO	EO	DR
7-6 1600	-0.03	-1.41	-1.81
7-6 2345	0.63	-1.17	-0.51
7-7 1050	0.45	-0.14	-0.01
7-7 1805	0.88	0.45	-0.56 *
7-7 2300	0.60	0.90	-0.47
7-8 0645	1.15	0.51	-0.44
7-8 1120	1.66	1.45	-0.35
Entire Mission	1.12	1.66	-0.11

	EO FIRST	DR FIRST
24 Hours as EO	-0.26	1.87

MISSION 2

DATE/TIME	RO	EO	DR
7-11 1515	-0.87	-0.55	-1.81
7-11 1830	-1.54	-1.44	-2.14
7-11 2315	-1.17	-0.69	-1.50
7-12 0545	-1.14	0.42	-0.63
7-12 1100	0.21	0.21	-0.51 *
7-12 1515	-0.36	-0.33	--
7-12 1745	0.99	0.09	-0.45
7-12 2230	0.27	2.35	-0.42
7-13 0645	0.12	1.48	-0.12
7-13 1000	-0.03	0.90	-0.36
Entire Mission	-0.01	0.96	0.42

	EO FIRST	DR FIRST
24 Hours as EO	-.87	1.29

* EO and DR change position

APPENDIX F

PROSPECTIVE WORKLOAD RATINGS OF LOS-F-H MOBILE AIR DEFENSE MISSILE SYSTEM

Susan G. Hill James C. Byers Allen L. Zaklad
Alvah C. Bittner, Jr. Richard E. Christ

Abstract

Prospective ratings of operator workload (OWL) were obtained from six operators of the Line-of-Sight-Forward-Heavy (LOS-F-H) air defense system. Using the Task Load Index (TLX), ratings of predicted workload were obtained for four separate topic areas: new equipment, multiple fire units, multiple targets, and crew organization. Analyses of variance of TLX global and subscale scores revealed significant differences between OWL ratings for current and proposed operation in the four topic areas. Use of ratings to prospectively estimate OWL of systems and events is discussed.

INTRODUCTION

The Line of Sight-Forward-Heavy or LOS-F-H is an air defense system with a requirement to engage low-altitude helicopters and fixed-wing threat aircraft. A Non-Developmental Item Candidate Evaluation (NDICE) was conducted in 1987 and the winning system was selected to be the "baseline" LOS-F-H. Initial operator workload (OWL) assessments of the winning candidate were conducted retrospectively, by asking the soldier-operators to make judgments of OWL after viewing videotapes of their own performance during NDICE (Hill, Zaklad, Bittner, Byers, & Christ, 1988) and to make overall judgments of various generic mission segments and tasks (Bittner, Byers, Hill, Zaklad, & Christ, 1989) --See respectively Appendices B and C of this report.

A Force Development Test and Experimentation (FDTE) program for the LOS-F-H system was held in June-July, 1988, at Fort Bliss, TX. The purpose of this field test was to examine selected concepts regarding tactics, doctrine, organization and training. The test took place over a six-week period, with the first five weeks comprised of one-hour missions and the last week including two 48-hour missions. The OWL assessments of various tasks under a variety of mission contexts for both the "basic" four-hour missions and the sustained 48-hour missions are described and discussed by Hill, Byers, Zaklad, and Christ, 1989a and 1989b, respectively -- see also Appendices E and D of this report. The present study is the fifth in this series of investigations. It builds upon the background of empirical OWL investigations by using OWL ratings as a basis for predicting the workload that will be associated with modifications in the system and its operational context.

Background

Workload has become an area of concern as technology advances and operator functions are increasingly cognitive in nature. (See Lysaght et al., 1988, for an integrative review of OWL literature.) Of particular interest are methods to estimate or predict OWL early in system development. One such method involves subjective ratings of workload made in conjunction with descriptions of systems or events that have not yet been personally experienced by the individuals making the ratings. These are referred to as prospective or projective OWL ratings.

Prospective ratings have been employed in several previous applications. Several early studies were performed using the Subjective Workload Assessment Technique or SWAT (Reid, Shingledecker, & Eggemeier, 1981) provided encouraging results (Eggleston, 1984; Eggleston & Quinn, 1984; Reid, Shingledecker, Hockenberger, &

This appendix contains a revised and condensed version of unpublished Technical Memorandum Number 2, prepared by the indicated authors.

Quinn, 1984). More recently, Masline and Biers (1987) compared projective subjective workload assessments of a task which had been described in written and verbal form to assessments of the same task experimentally performed. The subjective assessments were obtained via three psychometric scaling techniques (magnitude estimation, equal appearing intervals, and SWAT). Results suggest that subjects gave similar workload assessments whether they did so projectively or actually performed the task. Masline and Biers do caution that insufficient research has yet been done to make any generalizations about the validity of prospective workload assessments. The results so far are promising and further research is clearly warranted.

Purpose

The research presented in this paper has two objectives: (a) to examine the use of OWL rating scales to obtain prospective estimates of workload, and (b) to provide prospective estimates of OWL that may be used in LOS-F-H system development.

METHOD

Prospective OWL ratings were obtained at the conclusion of the FDTE field exercises. The availability of the LOS-F-H operators during this period made the present study possible. In addition, the FDTE had provided training for the operators in both system operation and judgments of operator workload using the Task Load Index (TLX) (Hart & Staveland, 1987). A final rationale for using the FDTE as the context for this study was the upcoming FDTE-Phase II which was scheduled for the summer of 1989. The prospective OWL measures administered during the initial FDTE could be later validated with actual data obtained during a Phase II FDTE.

Subjects

The subjects were six soldier-operators who had been participants during both the NDICE and FDTE tests. The operators included two radar operators (ROs) and four electro-optical operators (EOs). The ROs also served as squad leader and mission commander; the EOs also served as gunners. The EOs were junior enlisted men (Private First Class and Specialist) and the ROs were junior Non-Commissioned Officers (Sergeant).

Workload Scale

The TLX scale was used to collect ratings of OWL. The TLX is a multidimensional scale composed of six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration, each rated on a scale from 0 to 100. A weighting procedure is used to combine the six individual subscale ratings into a global or composite workload score. Normally, each rater will designate, for each task to be rated, the more important of all possible pairs of the six subscales. For this study of prospective workload rating the standard procedure for determining weights was not followed. This deviation from standards was deemed necessary because the tasks and the conditions in which the tasks were to be performed had never been experienced by the rater. Instead, all the TLX scores used for the present study were weighted by each soldier's paired comparison weights for the "Entire Acquisition/Tracking Mission," as they were originally obtained for the workload analysis of basic four-hour missions in the FDTE (see Hill el al., 1989b or Appendix D of this report).

An advantage of a multidimensional scale such as the TLX is that it provides the ability to look at the separate subscales for diagnostic analysis. Other reasons for choosing TLX are that experience had shown that it could be quickly completed, it was well accepted by the soldiers, and it had demonstrated consistently higher validities when used for direct assessment (see, for example, Byers, Bittner, Hill, Zaklad, & Christ, 1988 and Hill et al., 1989a and 1989b).

Topic Areas

Four distinct topic areas were chosen for prospective investigation using the TLX rating scales. These were new equipment, multiple fire units, multiple targets, and crew organization. New equipment and crew organization represent optional system modifications, whereas multiple fire units and multiple targets reflect a more realistic tactical context.

New equipment. This topic area refers specifically to automated radar. It includes automated identification of blips as targets; automated identification of the target as fixed- or rotary-wing; and automated prioritization of targets, with appropriate symbology displayed on the radar display. Even with automated radar, however, the

RO would continue to monitor the radar and make decisions as necessary (e.g., change priorities based on other information). The subjects were asked to make prospective ratings of the workload for the RO and EO using this new radar equipment.

Multiple fire units. This topic area represents a change from the FDTE condition. It refers to a configuration of a master fire unit controlling one or more slave fire units. It assumes some form of automated radar (as described in the previous paragraph). The master fire unit radiates radar signals, receives Command and Control (C2) data, and determines the assignment of targets to fire units in the platoon. The slave fire unit receives target information via a local C2 communication channel, is responsible for the target assigned, and searches for other targets of opportunity. The soldiers were asked to make prospective ratings of the workload for the RO and the EO in the master unit and for those in a slave unit.

Multiple targets. This situation refers to the case in which more than a single target appears at one time. The first set of OWL ratings asked the soldiers to rate RO and EO workload for double the number of targets that they had been seeing during a one-hour acquisition/tracking mission segment in the FDTE. A second set of OWL ratings asked for RO and EO workload in the situation where two fixed-wing aircraft (in attack profile) and two pop-up helicopters appeared in rapid succession. The concern here was that the serial nature of the RO and EO tasks in an engagement sequence leads to easy handling of single targets, but to potential problems when many targets rapidly appear.

Crew organization. At the time of this study, the LOS-F-H had a crew of three: the RO, the EO, and the driver (DR). The RO monitors the radar to analyze, plan, and conduct the air battle. However, the RO must also function as the squad leader and mission commander (MC) for the fire unit, responsible for performing many C2 functions both for the fire unit and for the maneuver unit that is being supported. The EO is the gunner and has the primary job of tracking and engaging targets. The DR handles the vehicle, but otherwise has little to do. This crew organization was used during the NDICE and FDTE, both of which involved a single fire unit with no maneuver unit to support or other asset to protect, and with little communication and cross-country navigating.

Because the DR has little to do, there has been some discussion of a reorganization of the crew to more equally distribute workload. Furthermore, there was some concern that the RO/MC could not adequately perform many of the functions required of that position in a realistic battlefield scenario. A proposed crew organization included suggestions which would change the physical location, duties, and responsibilities of some crew members. In this reorganization, the senior ranking MC would occupy the DR's position, from which he would keep the fire unit in the battle and monitor the ground battle. DR/MC would also maintain direct contact with the platoon leader, have visual contact and voice communication with the maneuver force or asset, drive the vehicle and serve as the "eyes" for the RO and EO. The RO, under this reorganization, would coordinate the tactical air battle and respond to an integrated weapons display for analysis, operation, and planning. The EO would continue to conduct engagements and serve as the backup for the RO. Essentially, in the proposed organization, the MC no longer functions as the RO but instead as the DR.

Soldiers were asked to rate easy and difficult missions for each of three crew positions with current organization and job requirements (i.e., RO/MC, EO, and DR) and with the new proposed organization and job requirements (i.e., RO, EO, and DR/MC). Easy missions were characterized by day operations in a shirt-sleeve environment, with no smoke or little electronic countermeasures (ECM). Difficult missions were described by day operations in full chemical protective gear, heavy ECM, and many targets.

Procedure

The prospective workload ratings were obtained during the sixth and seventh weeks of FDTE testing. While one crew was participating in its 48-hour mission, the other, "off" crew, performed the prospective OWL ratings. Hence, the two crews participated in the prospective ratings under somewhat different conditions, at different times, and in different test locations. Since the topic descriptions were given verbally, the two presentations of the same information may have differed slightly. In addition, one crew had not yet participated in its 48-hour mission, while the other had completed it when they did the prospective ratings. It is not believed that these differences had any significant effects on the ratings obtained.

The same procedure was followed for both crews. Upon arrival, the purpose of the session and the procedure to be used were explained. First, five OWL ratings of the FDTE 4-hour mission just completed were obtained: Overall FDTE, average day and average night missions in MOPP 0 and in MOPP 4. Then, the first prospective topic area given above was described and ratings were made by the crew. The completed ratings were collected and then the crew members were asked what they thought about the topic and its potential impact on the system and system operation. This procedure was repeated for all four topics area, in the order used earlier in this section.

A total of 27 OWL ratings were made by each of the six soldiers. Five concerned workload of the just completed FDTE. Twenty-two involved prospective workload ratings for the four topic areas described previously: two for new equipment, four for multiple targets, four for multiple fire units, and 12 for new organization.

RESULTS

For each topic area, comparisons between current situations and proposed future conditions were made. The results obtained for composite or global TLX scores are reported separately for each topic area in terms of their statistical significance, displayed graphically, and briefly described narratively. Although different opinions were expressed during the informal discussions concerning each topic area, a consensus was generally reached. The essence of these discussions is presented following the presentation of global workload data for each topic. The results for subscale ratings are presented separately, after the results for global scores and operator opinions.

New Equipment

An analysis was performed comparing workload ratings of automated radar to ratings of the current (non-automated) radar equipment for an average mission. For this analysis the ratings for current radar equipment for an average mission were derived by averaging ratings of easy and difficult missions. A three-way analysis of variance (ANOVA) was performed with factors of Radar Configuration (automated and current), Position (RO and EO), and Subscale (6 TLX dimensions). This analysis revealed a significant difference between Automated Radar and Current Radar ($F(1,5) = 7.30$, $p < 0.043$). The Automated Radar had lower workload ratings than the current configuration (21.7 and 31.7, respectively). The interaction between Radar Configuration and Position was also significant ($F(1,5) = 14.79$, $p < 0.012$); the RO experiences a somewhat greater reduction in OWL than the EO (32.5 to 19.2 and 31.2 to 24.2, respectively).

Soldier comments were consistent with these statistical results (e.g., "The automated radar would be nice to have. The RO wouldn't have much to do with the automated radar. It would be really helpful." "It would be like a previous system where the radar set up tracks, prioritized targets and everything.")

Multiple Fire Units

Analysis was performed comparing the Master and Slave Modes to Autonomous operation. For this analysis the ratings for Autonomous operation were derived by averaging ratings of easy and difficult mission (i.e., they were the same values as those used for current radar equipment above). Specifically, a three-way ANOVA was performed with factors of Mode (Master, Slave, or Autonomous), Position (RO and EO) and Subscale (6 TLX dimensions). This analysis revealed no main effect of position or mode. However, the Mode-by-Position interaction was significant ($F(2,10) = 18.20$, $p < 0.0005$). As shown in Figure F-1, the total workload for RO and EO is rated about the same in the autonomous mode. However, this figure also shows that the RO is judged to have much greater workload than the EO in the Master Mode, and, conversely, the EO is judged to have greater workload than the RO in the Slave Mode.

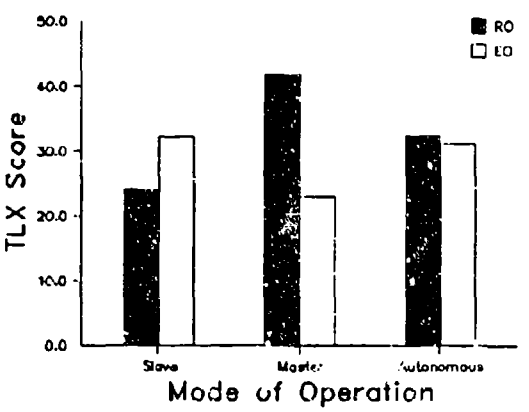


Figure F-1. *The effect of crew member position and mode of operation on prospective TLX ratings.*

Soldier comments were generally consisted with the ANOVA (i.e., "What is the RO in the slave going to do?" "The RO in the master would be really busy." "The EO really wouldn't change in the slave from what it is now.")

Multiple Targets

The two multiple target situations were examined separately. First, an analysis was performed to test the differences between workload ratings for double the number of targets and for the average mission. Specifically, an ANOVA compared the Mission Target Density (Double and Average), Position (RO and EO) and Subscale (6 TLX dimensions). Mission Target Density was revealed by this analysis to have a significant main effect ($F(1,5) = 9.26$, $p < 0.03$). Double Targets had a mean global TLX rating of 46.2, while the average mission had a workload rating of 38.7. There were no significant interactions.

An ANOVA comparing the workload rating of a two fixed-wing and two-rotary wing (2FW2RW) pass and an average mission, Position (RO and EO), and Subscale (6 TLX dimensions) was performed. There was a significant difference in mean TLX workload ratings for 2FW2RW and Average ($F(1,5) = 16.50$, $p < 0.01$), with the means being 45.8 for 2FW2RW and 31.7 for average mission workload. As in the double target configuration, there were no interactions.

Soldier comments were in line with the quantitative results (i.e., "With more targets, it would be pretty busy." "With the two fixed wing aircraft and two pop-up helicopters, the crew might not be able to get them all." " More helicopters, such as five popups, would be the toughest situation.").

Crew Organization

ANOVA was performed to compare Current and Proposed Crew Organization, Mission Difficulty (Easy and Hard), Position (RO, EO, and DR), and Subscale (6 TLX dimensions). A significant main effect confirmed that OWL ratings were greater for hard missions than easy missions ($F(1,5) = 17.12$, $p < 0.01$). The mean TLX workload rating for easy missions is 21.0 while that for difficult missions is 39.1.

A significant interaction between Organization type and Position was also found ($F(2,10) = 4.57$, $p < 0.04$). In the current organization, the RO/MC and EO have about equivalent OWL ratings while the DR has much less workload (31.7, 30.8, and 19.1, respectively). In the proposed organization, all three positions have similar workload (i.e., the OWL is leveled across positions). For the RO, EO, and DR/MC, mean TLX ratings were 33.5. 31.0, and 33.4, respectively.

Figure F-2 shows the interaction among Position, Organization, and Mission Difficulty. Although not statistically significant ($p < 0.12$), the data suggest that the proposed organization would be most beneficial for more difficult missions. Thus, not only is there a more equitably distributed workload across crew positions in the difficult mission condition but there is also a reduction in the absolute amount of workload for both the RO and EO when they are most likely to need some unburdening.

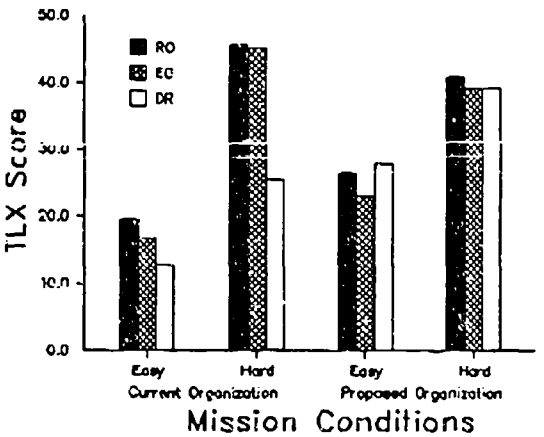


Figure F-2. *The effect on prospective TLX ratings of crew member position, mission difficulty, and crew organization.*

Soldiers commented that the proposed organization sounded very strange. One current squad leader said he didn't mind the idea of driving and said he'd like to be able to see out of the vehicle and see where he was. Currently, he stops the vehicle at times and gets out so he can look around. The other squad leader does not want to be the DR. He drove for someone else and now that he's promoted, he wants someone to drive him around. The two EOs in this latter crew don't want the squad leader to be the DR because they are looking to promotion and want somebody to drive them. Soldiers' comments reflected current views as to the status of driving.

Subscale Results

The main effect of subscale was significant in each of the five ANOVA in which it was used as a source of variance: for New Equipment, Multiple Fire Units, Double Targets, 2FW2RW Targets, and Crew Organization, $F(5,25)$ = 2.95, 2.89, 3.01, 2.99, and 3.75, respectively, all with $p < .03$). In order of decreasing magnitude, the mean weighted subscales scores averaged over all five sets of data are as follows: Mental Demand (142), Temporal Demand (98), Effort (95), Performance (94), Frustration (57), and Physical Demand (12). There were no significant interactions involving subscale in the ANOVAs applied to data for the new equipment, the two multiple target conditions, or the crew organization. For the multiple fire unit data there were significant interactions for Mode and Subscale ($F(10,50)$ = 2.74, $p < 0.009$), and for Mode, Position, and Subscale ($F(10,50)$ = 3.66, $p < 0.001$). The two-way interaction is driven principally by the fact that both Mental and Temporal Demands are less in the Slave mode (112 and 74, respectively) than in either the Master (153 and 103) or the Autonomous (131 and 88) Modes of Operation.

The three-way interaction involving Mode, Position, and Subscale is illustrated in Figure F-3. This figure shows that the smaller level of Mental and Temporal Demands for the Slave Mode of Operation, noted in the Mode-by-Subscale interaction, are primarily due to the Slave-RO ratings being substantially lower than those for the RO in the Master and Autonomous conditions. Another major 3-way trend in the data shown in Figure F-3 is that (a) both Mental and Temporal Demands are higher for the RO than the EO in the Master Mode, (b) both are lower for the RO than the EO in the Slave mode, and (c) they are essentially equal for the RO and EO in the Autonomous Mode. There is the appearance of other effects as well, such as the extraordinarily high level for Effort for the RO in the Master Mode in comparison to all other combinations of Position and Mode of Operation.

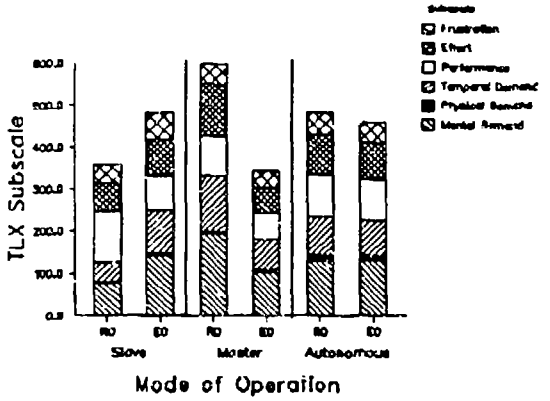


Figure F-3. *The effect on prospective TLX ratings of TLX subscale, crew member position, and mode of operation.*

DISCUSSION

This investigation jointly evaluated the prospective use of the OWL scales and some aspects of workload for the LOS-F-H. It represents the first in a programmatic series of empirical investigations aimed at prospective estimations of OWL in Army systems. Discussed in succeeding sections are prospective OWL assessments organized by topic areas, use of prospective assessments, and future work.

Prospective OWL Assessments for Four Topic Areas

The four topic areas produced different overall levels of workload ratings, with multiple targets yielding the highest ratings and current "easy" missions yielding the lowest.

New automated radar equipment. The soldiers clearly thought the automated radar would entail much less workload for the RO than the current system. This was apparent in both the ratings and informal discussion. Interestingly, the ratings further suggest that the soldiers felt the EO would have less workload as well (though a smaller reduction than the RO). This perhaps was due to the perception that improved processing of potential targets by the RO will lead to smoother and quicker handoff to the EO, thereby allowing the EO to perform his job with less workload.

Multiple fire units. The multiple fire unit situation does not represent an optional system modification, but rather a closer approximation to a realistic battlefield situation. Thus, the question is not whether to implement the modification or not, but how best to deal with the associated problems. From the global workload rating data, it is apparent that there is a potential function allocation problem (see Figure F-1). Any disparities in OWL levels between RO and EO for autonomous (or, as assessed here, average) missions will probably become exacerbated as the missions

get more difficult (i.e., as realism increases). Crew reorganization appears in one way to ameliorate this potential increase in both absolute levels and the variance of workload.

Multiple targets. There was significantly more workload judged both for double the number of targets in a one-hour target engagement period and for the 2FW2RW threat pass than for the average mission. The total amount of workload was judged to be about the same for both types of multiple target conditions (respectively, 46.1 and 45.8 on a 0-100 scale). An interesting methodological issue involves the OWL ratings for brief intervals (2FW2RW) and those for extended periods (double targets). The soldiers in this study were able to make both kinds of OWL ratings, but comparison of OWL ratings over different time periods leads to some logical difficulties.

Crew organization. Before conducting the formal ANOVA, it was suspected that the proposed reorganization would have an overall benefit for difficult missions. Such a benefit, it was suspected, would occur because of the redistribution of the increased workload (due to the mission difficulty) more evenly among the 3-man crew. Such an effect would be manifested in a significant interaction: Position X Mission Difficulty X Organization. Figure F-2 suggests such an interaction, but it is nonsignificant ($p < 0.12$). The lack of significance of this interaction may be partially due to the soldiers' inability to assess the impact of the reorganization. This topic area was the least familiar to the soldiers.

A final point before leaving the issue of topic area is that we would anticipate substantial interaction effects on workload by the joint impact of changes in all four of these topic areas. For example, it may be the case that advantages in the proposed crew organization would become most evident with the addition of coordination tasks (multiple units) and more difficult missions but that improved radar (and other new) equipment would somewhat negate the need for the new organization.

Subscale Analysis

The significant main effects of the subscales for all five analyses showed that there are differing dimensions contributing to a perception of workload. Clearly, for the system under study, physical demand contributes the least while mental demand contributes the most to the perception of workload. Based on the tasks required to successfully operate the LOS-F-H, and more pertinently, to engage targets with the LOS-F-H, these results are not at all surprising. These tasks are primarily cognitive and perceptual (there is relatively little manual or psychomotor activity). A remaining question concerns the other four dimensions of the TLX scale: temporal, performance, effort, and frustration. In general, the first three are usually close together and greater than the frustration rating. However, the presence of significant interactions of both crew position and operational mode with subscales, as well as other trends in these data, beg that more work be done to sort out the impact of these dimensions on the overall experience of workload.

Examination of the diagnosticity of the TLX subscales requires more detailed analyses than is within the scope of the present study. However, the ability to examine workload ratings in a finer level of detail can be seen to be a major advantage of multidimensional scales such as NASA TLX.

Using OWL Scales for Prospective Assessments

Several observations can be made regarding the use of TLX to obtain prospective OWL ratings. One observation was that the soldiers did not appear to be comfortable passing judgment on potential changes, and the impact of changes, in the air defense system under study. This was perhaps due to the newness and the developmental status of the LOS-F-H. The crew members were least hesitant to pass judgment on topics for which they had some previous relevant experiences. In this observation, there is some suggestion that in order to successfully apply prospective techniques, the subjects must have some experience relevant to the topic in question. The one topic area that did not have such a basis for comparison -- proposed crew organization -- produced problematical results, possibly due to the absence of a relevant "comparative" anchor. It might also be that insufficient detail was given to the subjects concerning the proposed modifications. Consequently, in the areas in which the soldiers had some prior experience, they perhaps filled in detail themselves, while in the topic area in which they had no experience, they were unable to fill in sufficient detail. In either case, it seems clear that the prospective techniques cannot be used on topics that are "completely out of the blue."

A second observation is concerned with the weightings used to reflect the importance of the various subscales. For prospective ratings, which weightings should be obtained and used? Should the weightings be made prospectively as well as the ratings? The decision made for the present investigation was to use the weightings that had been previously obtained for engagement missions. It was felt that the prospective ratings were a sufficient challenge to the soldiers and asking them to make further future projections would not necessarily add information. The engagement mission weightings would give weightings that reflected the individual importance of the various subscales to the perception of workload while accomplishing the engagement mission. More thought should be given to the question of what are the must appropriate TLX weightings to be used in a prospective application.

The prospective workload ratings obtained in this study were average ratings for generic mission segments and tasks; they are not fine-grained ratings reflecting the impact of detailed information on mission conditions (see Bittner et al., 1989). However, it would be interesting to have prospective ratings made for very precisely defined mission scenarios. Much more information of great value for predicting workload in potential future circumstances could be examined if this were to be accomplished. Comparisons with potential individual and system-level performance would also be possible. The use of rapidly reconfigurable interactive soldier-in-the-loop simulators might be desirable to achieve this objective.

Topic descriptions were given verbally in this study. Although this seemed to work successfully, it is possible that written descriptions would give more assurance that all subjects were making workload ratings of the same event. Perhaps better still would be the use of soldier-in-the-loop simulators to convey a "common" sense of a future system configuration to the raters. This is an area for future work.

Future Work

The next step in this research would be to compare these prospective ratings with empirical ratings of the same modifications or mission events. Examining how empirically-collected OWL ratings correspond to the prospective would serve to validate the prospective. Experiencing how both the prospective and empirical subjective measures relate to system performance measures would also be of interest. A problem that may be anticipated is that of matching the topic descriptions given in the prospective study with actual events in any simulation or test environment. To address this, criteria should be developed prior to any "matching" of events so that only those events which satisfy the criteria may be used. However, even after meeting these criteria, any actual event will contain mission-specific occurrences not addressed in the prospective description. The question consequently will be whether judgments are being made of comparable events.

A second problem is concerned with the subjects used in the empirical data collection. It is uncertain that soldiers who participated in this study will be participating in future system testing. How appropriate is it to compare results obtained from a prospective and a real-time application of workload ratings if the two sets of ratings are made by different raters? If the same soldiers participate, will intervening experience have made comparisons between the prospective and actual OWL ratings incomparable? In any case, training and experience with the rating scale must be at a high level if the ratings are to be stable, as was true with the subjects used in the present prospective study.

An attempt to validate the prospective OWL ratings obtained in this investigation with empirical OWL and system performance data on the same system in the Phase II FDTE would have been well worth the effort even with these problems. Methodologies to predict operator workload early in the design and development of system and organizational concepts are critical to optimizing future forces.

CONCLUSIONS

Three conclusions may be drawn from the present evaluation of the use of an OWL scale in prospective workload assessments:

(1) TLX may be used by soldiers to make OWL ratings of events that had not yet been experienced. Soldiers felt they were making meaningful judgments of workload for the verbally described situations.

(2) The prospective ratings have face validity (i.e., ratings made sense and reflected what might be expected). However, these results must be

compared to empirical and performance data collected in the future for validation of the correspondence.

(3) Use of subscale data from multi-dimensional workload techniques is of potential diagnostic value and warrants further evaluation (e.g., TLX, SWAT).

It is too early to suggest that the prospective assessment is a valid and reliable method for predicting system OWL. More research regarding validation of prospective OWL ratings needs to be conducted. There is the need for application of such prospective techniques to actual system design and development, where predictive estimates may be compared to the empirical. In addition, it would be of considerable interest to compare prospective and empirical measures with operator and system measures of performance. How the predicted estimates of workload from the prospective methodology are associated with the results of other analytical or empirical OWL measures also is an area for future investigation.

REFERENCES

Bittner, A. C., Jr., Byers, J. C., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1989). Generic workload ratings of a mobile air defense sSystem (LOS-F-H). Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1476-1480). Santa Monica, CA: Human Factors Society.

Byers, J. C., Bittner, A. C., Jr., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1145-1149). Santa Monica, CA: Human Factors Society.

Eggleston, R. G. (1984). A comparison of projected and measured workload ratings using the subjective workload assessment technique (SWAT). Proceedings of the IEEE National Aerospace and Electronics Conference (pp. 827-831). Dayton, OH

Eggleston, R. G., & Quinn, T. J. (1984). A preliminary evaluation of a projective workload assessment procedure. Proceedings of the Human Factors Society 28th Annual Meeting (pp. 695-699). Santa Monica, CA: Human Factors Society.

Hart, S. G., & Staveland, L. E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.S. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Hill, S. G., Byers, J. C., Zaklad, A. L., & Christ, R. E. (1989a). Subjective workload assessment during 48 continuous hours of operations of the LOS-F-H. Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1129-1133). Santa Monica, CA: Human Factors Society.

Hill, S. G., Byers, J. C., Zaklad, A. L., & Christ, R. E. (1989b). Subjective workload ratings of the LOS-F-H mobile air defense missile system in a field test environment (Technical Memo 5). Willow Grove, PA: Analytics, Inc.

Hill, S. G., Zaklad, A. L., Bittner, A. C., Jr., Byers, J. C., & Christ, R. E. (1988). Workload assessment of a mobile air defense missile system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1068-1072). Santa Monica, CA: Human Factors Society.

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., Wherry, R. J., Jr., Zaklad, A. L., & Bittner, A. C., Jr. (1988). Operator workload: Comprehensive review and evaluation of workload methodologies (Technical Report 851). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Masline, P. J., & Biers, D.W. (1987). An examination of projective versus post-task subjective workload ratings for three psychometric scaling techniques. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 77-80). Santa Monica, CA: Human Factors Society.

Reid, G. B., Shingledecker, C. A., & Eggemeier, T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting (pp. 522-525). Santa Monica, CA: Human Factors Society.

Reid, G. B., Shingledecker, C. A., Hockenberger, R. L., & Quinn, T. J. (1984). A projective application of the subjective workload assessment technique. Proceedings of the IEEE National Aerospace and Electronics Conference (NAECON) (pp. 824-826). Dayton, OH.

DATA ATTACHMENT F-1

Factor Scores for the LOS-H-H FDTE 48-Hour Mission Study

		Average Day Mission		Average Night Mission		Automatic Radar	
	Overall	MOPP 0	MOPP 4	MOPP 0	MOPP 4	RO	EO
RO #1	50.4	37.7	43.7	42.3	38.3	20.3	40.3
RO #2	52.9	36.0	42.0	23.3	45.7	19.3	16.0
EO #1	25.4	7.7	8.0	7.7	8.3	3.7	8.0
EO #2	28.1	20.3	20.3	13.3	22.3	11.7	14.0
EO #3	59.4	24.0	41.0	13.7	54.0	22.7	22.7
EO #4	56.0	46.3	56.7	42.7	56.3	37.3	44.3

	Slave System		Master System		Double Targets	
	RO	EO	RO	EO	RO	EO
RO #1	33.7	39.7	53.7	30.3	35.7	45.7
RO #2	14.3	28.7	41.3	20.7	42.0	25.0
EO #1	3.7	9.0	12.7	7.7	13.7	16.3
EO #2	6.3	14.0	20.3	16.0	46.3	50.7
EO #3	49.0	51.3	68.0	28.3	81.0	64.7
EO #4	37.7	52.0	56.3	36.0	64.7	68.3

	2FW+2RW Targets		Average Mission	
	RO	EO	RO	EO
RO #1	54.3	49.0	29.3	27.7
RO #2	44.7	27.7	33.5	32.7
EO #1	13.7	22.7	14.0	12.3
EO #2	28.0	28.0	20.2	20.5
EO #3	75.0	69.0	53.2	48.8
EO #4	67.0	70.7	45.0	45.0

DATA ATTACHMENT F-1 (Continued)

	Easy Mission			Hard Mission		
	RO	EO	DR	RO	EO	DR
Missions With Current Crew Organization						
RO #1	23.7	20.3	28.3	35.0	48.3	63.0
RO #2	11.0	9.3	5.0	56.0	45.3	10.0
EO #1	8.7	5.3	3.7	19.3	21.0	1.0
EO #2	11.3	12.0	15.0	29.0	22.7	10.0
EO #3	25.3	16.7	7.0	81.0	75.3	47.3
EO #4	37.0	37.0	18.0	53.0	57.0	21.0
Missions With Proposed Crew Organization						
RO #1	41.7	26.3	35.7	35.0	27.7	47.0
RO #2	30.3	14.3	27.0	55.7	45.7	49.7
EO #1	9.7	8.0	15.3	18.0	17.7	17.0
EO #2	22.0	26.3	30.0	4.1	29.7	34.3
EO #3	19.7	31.7	19.7	30.3	50.7	38.7
EO #4	34.0	31.0	38.7	65.0	63.0	48.0

APPENDIX G

WORKLOAD ASSESSMENT OF A REMOTELY PILOTED VEHICLE (RPV) SYSTEM*

James C. Byers Aivah C. Bittner, Jr.
Susan G. Hill Allen L. Zaklad Richard E. Christ

Abstract

Four empirical operator workload (OWL) scales were applied to ground control operations of the Aquila remotely piloted vehicle (RPV) during a recent field test: Task Load Index (TLX), Subjective Workload Assessment Technique (SWAT), Overall Workload (OW), and the Modified Cooper-Harper (MCH). Seventeen sets of individual assessments of mission segments were made by the four members of each of four crews and one replacement crewman. Jackknife factor analysis revealed the presence of only a single factor and indicated that the mean factor loadings formed a consistent ordering ($F(3,48)=503.5$, $p<.0005$): TLX (.910), SWAT (.893), OW (.869), and MCH (.833), with all pair-wise differences significant. Analyses of variance also examined the effects of test variables on the composite workload factor scores; significant findings were found which reflected both on the system and its test. These findings as well as informal lessons learned are discussed in the context of the development and validation of a methodology for assessing OWL.

INTRODUCTION

Four operator workload (OWL) scales were administered to Aquila remotely piloted vehicle (RPV) ground control station (GCS) crew members as part of a field test conducted during the period from October through November, 1987. The field test, run as part of a Force Development and Test and Experimentation (FDTE) program, was aimed at examining operational and organizational issues, particularly those associated with the ability of the GCS crew to plan and execute a simulated RPV reconnaissance mission. It was clear that target detection performance was the principal concern of the FDTE and that nothing would be allowed to interfere with obtaining optimal performance of the system. There was also the sense that the fate of the Aquila system depended on the soldiers' performance during the FDTE.

Background

A major deficiency discovered in the RPV system during an earlier Operational Test (OT) II was the inability of the GCS crews to satisfactorily detect, recognize, and locate target arrays. The target acquisition deficiency was considered to be solvable. New software programs were developed to support new automated search routines and to calculate and control various flight parameters. New hardware was developed which would create a compressed time plot of the mission for planning purposes. The personnel assigned to the GCS were given additional training designed to improve their ability to perform.

In addition, to improve the ability of the crew to "negotiate" mission parameters, to plan the mission, as well as to improve target acquisition performance, a fourth member was added to the crew, a Commissioned Officer (1LT or 2LT) with tactical knowledge and expertise. This Commissioned Officer would become the crew chief and mission commander (MC). The air vehicle operator (AVO) and mission payload operator (MPO) positions would remain the same as they were in the Aquila OT II (i.e., both positions were filled by enlisted personnel with the rank of private first class or specialist). The senior non-commissioned officer (NCO) or warrant officer who was previously the MC was now designated the RPV Technician (RPVT). However, the roles and relationships between the MC and RPVT were not clearly defined.

---

* This appendix contains a revised and condensed version of a paper presented at and published in the Proceedings of (pp. 1145-1149) the 32nd Annual Meeting of the Human Factors Society.

Since the major issue of the Aquila FDTE was target acquisition, system performance factors largely controlled by the MPO, the Aquila mission payload package (i.e., the camera, communication, and designator equipment) was mounted to the underside of a small, highly maneuverable aircraft. The pilot of the manned aircraft would respond appropriately to the inputs of the GCS computer and the AVO. This change from normal Aquila operational procedures would enhance the safe operations of the RPV. Also, since the mission payload package was mounted on a manned aircraft, the potential risk involved in launching and recovering the RPV was considerably reduced.

Purpose

As part of the FDTE, the present effort was concerned with workload variations across mission segments, crews, and crew duty positions as well as relative workload differences between the FDTE and the OT II. In addition to these system concerns, the present investigation was also concerned with the broader issues that concerned the relative efficacy and operator acceptance of four alternative OWL rating scales and of the applicability of the OWL scales under conditions characterizing field evaluations.

METHOD

Subjects. Operator ratings were obtained from 17 GCS crew members, four crews each consisting of a MC, AVO, MPO, and RPVT, and one replacement soldier. The MC was a lieutenant, the AVO and MPO were lower ranking, enlisted personnel, and the RPVT was a senior NCO or a warrant officer.

Procedure and instruments. Twenty-three separate Aquila RPV flights were used to conduct seven different sets of mission orders. These 23 flights were distributed over four 4-man crews, where one crew planned and flew five missions and three planned and flew six missions each. Each crew member made individual ratings of OWL during post-mission sessions for each mission which was planned and flow by his crew. Two segments of each mission were rated for at least four missions: Mission Planning and Flight. Eight other mission segments (e.g., detecting stationary versus moving targets) were also rated in one or more missions but they were not consistently rated due, in part, to the constrained conditions under which the data were being collected.*

Four workload rating scales were selected for evaluation in this study. These were the Task Load Index (TLX) (Hart & Staveland, 1987), Subjective Workload Assessment Technique (SWAT) (Reid, Shingledecker, & Eggemeier, 1981), Modified Cooper-Harper (MCH) scale (Wierwille & Casali, 1983), and Overall Workload (OW) scale (Vidulich & Tsang, 1987). These four scales were administered in counter balanced order over successive missions, crews, and crew members.

After the crew members had rated and discussed with the OWL team their experiences during the last mission they flew in the FDTE, those subjects who had also participated as GCS crew members during the OT II several months earlier were asked to use only the TLX and OW rating scales to make some additional workload ratings. These subjects (nine in total over all crews) were asked to provide average workload ratings for three mission segments encountered (though not necessarily rated for workload) during the FDTE. The mission segments of interest were: Mission Planning, Flight, and Target Detection. These nine subjects also were asked to recall their experiences during the OT II and to provide overall ratings for the same three mission segments as they were experienced during performance in the OT II.

Finally, subsequent to the assessment of overall workload in the FDTE and OT II, a rating scale questionnaire was administered to all 17 GCS participants. This questionnaire solicited judgments regarding the procedures and test instruments, particularly those used to measure OWL. The questionnaire asked the subjects to rate the four OWL instruments regarding: (a) Which they liked best; (b) which was the easiest to complete,

* For a number of reasons, the OWL data collection effort was forced to proceed under very constrained conditions. The data collectors were not allowed in the test environment of the GCS and had no access to GCS crew members prior to or during the conduct of a given mission. The crew members were interviewed and debriefed by FDTE test personnel following the completion of a mission, then transported to a separate facility in which they were administered workload assessments and interviews. Most constraining was the fact that the OWL data collectors were given limited or no advanced information about the test conditions which were to be employed during a particular Aquila mission. Consequently, the data collectors could not adequately prepare and key the OWL rating scales to specific types of mission segments prior to the arrival of the test subjects.

(c) which was the hardest to compiete; and (d) which allowed the best description (rating) of the workload that had been experienced. The administration of this questionnaire facilitated an open discussion of the four workload assessment scales.

RESULTS

Analyses were conducted in three phases which respectively examined: (a) the factor validities of the four workload scales; (b) an analysis of the workload associated with various test conditions; and (c) the summary results of the rating scale questionnaire.

Factor Validity Analyses

The analysis of factor validities was conducted in two stages. During the first stage, Principal Component Analysis (PCA) was conducted on the 349 sets of mission segment ratings collected across all subjects and missions during the FDTE (cf., Dixon, 1983). Each set included global workload ratings using the four scales. This analysis revealed a single component, hereafter called the **OWL factor**, which explained 75.2 percent of the total variance (the second eigenvalue was only 0.46). This analysis also yielded OWL factor scores which were the basis for the workload analyses reported in the next section. The results of this initial analysis supported the view that the four workload scales essentially provide assessments of a single common factor.

Jackknife PCAs were conducted of the workload measures during the second stage of the factor validity analysis to evaluate the stability of the factor loadings of the four scales (i.e., the correlations of each scale rating with the OWL factor). Jackknife analysis generally involves successive analyses (PCAs in the present case) dropping subjects one-at-a-time from the data set in order to provide an analysis of the stability of parameters estimates (Hinkley, 1983). In the present case, with four factor loadings and all 17 subjects, a 4 (loading) by 17 (subjects dropped) matrix was produced which could be analyzed by a conventional repeated measures analysis of variance (ANOVA). This ANOVA (Dixon, 1983) revealed a significant difference among the workload scale factor loadings ($F(3,48) = 503.5$, $p < 0.00005$). Subsequent analysis indicated a consistent ordering of the mean factor loadings:

TLX(.910), SWAT(.893), OW(.869), MCH(.833).

While pair-wise differences were all statistically significant, they may be negligible in practical terms.

Workload Analyses

Two ANOVAs were conducted examining the effects of various variables based upon the **OWL factor scores** which resulted as part of the earlier described overall PCA. These ANOVAs respectively focused on comparisons within the FDTE and comparisons between the FDTE and OT II.

Comparisons within the FDTE. An ANOVA was initially used to evaluate the effects of Crews (1, 2, 3, & 4) and Positions (MC, AVO, MPO, & RPVT) on OWL factor score ratings across missions (1 to 4) and Mission Segments (Planning & Flight). (The raw data for this ANOVA are given in Data Attachment G-1 of this appendix.) This analysis, enhanced with the "analysis of error variances" (Bittner & Morrissey, 1988) revealed significant effects for Position ($F(3,9) = 2.77$, $p = .05$); the Crew-by-Position interaction ($F(9,27) = 14.75$, $p < .0001$); and Mission Segment ($F(1,9) = 7.25$, $p < .025$).

The mean OWL factor scores for the MC and MPO positions (0.26 and 0.50, respectively) are higher than those for the AVO and RPVT positions (-0.69 and -0.34, respectively), but there is no difference between the mean levels of workload experienced in the MC and MPO positions or in the AVO and RPVT positions. However, the interaction effect shows that there is considerable individual variation in workload ratings for each particular position. This interaction effect may reflect different interactive styles of the four crews. For example, all four crew members in one crew (the one labelled "A" in Data Attachment G-1) had below average OWL factor scores. This crew was observed by the OWL team and others as having a "laid-back" attitude toward their performance.

The main effect of mission segment is a result of the Flight segment being rated marginally higher in OWL than the Mission Planning segment (0.11 and -0.25, respectively).

Comparisons between FDTE and OT II. An ANOVA was applied for comparison of OWL factor scores computed from the data collected from nine subjects immediately after they

participated in the FDTE and one month after they participated in OT II. (The raw data for this comparison are given in Data Attachment G-2 of this appendix.) Using two groups counter-balanced with respect to order of field test rated (FDTE or OT II first), this analysis focused upon overall differences between field tests (FDTE and OT II) and their constituent mission segments (Planning, Flight, and Target Detection). This analysis revealed a significant effect for Field Test ($F(1, 7) = 8.34$, $p < .025$) and Mission Segments ($F(2,14) = 4.05$, $p < .05$), as illustrated in Figure G-1.

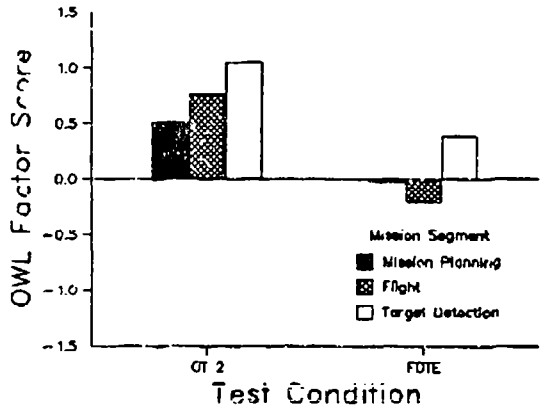


Figure The effect of mission segment and test conc on workload ratings.

Examining the figure, it may be seen that the mean O L factor scores associated with participating in the OT II were higher than those for the FD (overall, 0.78 and 0.06, respectively). Over both tests it may also be seen that the mean OWL factor score for Target Detection (overall, 0.72) was higher than those for Flight or Mission Planning (0.26 and 0.28, respectively), which were not different from one another.

Rating Scale Questionnaire Summary

Table 1 summarizes the quantitative results obtained from the subjects when they were asked to identify OWL assessment techniques which possessed certain specific features. It may be seen that most subjects both liked the TLX scale the best and believed that it provided the best description of the workload they had experienced. Subsequent follow-up interviews revealed that many who thought TLX provided the best description of the workload they experienced, liked it best for that reason.

Table G-1

Operator Acceptance of Workload Rating Scales in the Aquila RPV FDTE Study

	Rating Scale		
TLX	OW	MCH	SWAT
Which of the questionnaires did you like the best?			
7	3	3	1
Which questionnaire was the easiest to fill out?			
3	4	0	0
Which questionnaire was the hardest to fill out?			
2	0	8	2
Which questionnaire do you think best allowed you to describe the workload you experienced?			
10	5	2	0

Note. Data shown are the number of times each scale is given the highest ranking.

Regarding the relative ease and difficulty of using the different rating scales, most subjects thought the OW scale was the least difficult to complete and almost all indicated that the MCH scale was the hardest to complete. Follow-up interviews with the GCS crews revealed that the ease of completing a scale led some subjects to judge the OW scale as allowing the best description of workload. Not solicited from the subjects, but freely offered by most, were complaints regarding the difficulty of the SWAT card sort procedure which is required to scale workload ratings obtained with SWAT.

These results tend to indicate that operator acceptance is highest for the TLX assessment technique and lowest for MCH assessment technique within the limited subject group and conditions of the present investigation.

DISCUSSION

This investigation evaluated the use of four alternative OWL rating scales under field test conditions and the workload associated with

operating the GCS of the Aquila RPV. The results obtained for these two efforts are discussed in succeeding sections.

OWL Scales Under Field Test Conditions

This study demonstrated the successful application of a family of OWL assessment techniques in a stringent field test environment. The application for each of the techniques was under constraints much more severe than for most previous uses of the techniques, but not uncommon in field tests of interest to the Army. This application of OWL measures yielded formal and informal guidance regarding the use of these scales in field conditions.

Formal guidance. An ordering of the factor validities of the four measures was demonstrated during this investigation (TLX > SWAT > OW > MCH). In this ordering, little practical significance would be seen between TLX and SWAT; both of these have distinctly higher validities than OW and MCH. Between TLX and SWAT, however, the Ratings Questionnaire as well as complaints about the SWAT card sort procedure indicate that TLX was both: (a) more acceptable to most subjects and (b) believed to provide the basis for a better description of the workload that had been experienced.

Informal guidance. Much practical experience was gained concerning the assessment of workload during this FDTE. Several lessons learned are noted here:

- The initial briefing, separate from the post-mission data collection, was a convenient time to introduce the data collection team, the concept of workload, and the workload assessment scales. This initial briefing did entail coordination prior to test start in order to ensure the presence of all subjects;

- Providing refreshments (soft drinks and chips) to the crew members during post-mission data collection served several useful purposes. It staved off hunger so the crew members were willing to spend a little more time and thought on the assessment tools. More importantly, it provided a congenial atmosphere that helped to establish rapport; and

- The importance of talking with the crew members to obtain their impressions of what they do and why was confirmed during the test. Informal discussions with these subjects can give added insight into potential workload and other human factors problems.

Aquila GCS Workload

The workload analyses indicated significant effects for Crew Member Position, Mission Segments, and the interaction between Crews and Crew Member Position. In addition to confirming several anticipated findings, these results quantitatively supported observations of the workload assessment team. For example, the main effect for Position can be given the following interpretation. The generally higher ratings of the MCs is due to the fact that they were relatively inexperienced on the system and bore the responsibility for maintaining maximum levels of crew performance during a high visibility test. The workload experienced by the MPO was high since the focus of the FDTE was on target acquisition, the primary concern of the MPO. The lower workload of the AVO -- whose primary duty is to fly the RPV -- could be attributed to the fact that the RPV was not being flown; the mission payload package was mounted beneath a manned aircraft. The lower workload ratings of the RPVT reflect the ill-defined and non-relevant role they had in GCS operations during the FDTE, especially after serving as MCs during previous tests.

Discussions with crew members provide possible explanations for some of the results. For example, it was found that workload for flight segments of a mission was only marginally higher than that for planning the mission. Discussions with members of the crews suggest that much of the workload reported for mission planning resulted from the test situation and not from any intrinsic difficulty in mission planning.

The substantial difference in overall workload ratings between the FDTE and the OT II has several possible explanations. This difference in the experience of workload may reflect the more inclusive scope of the OT II when compared to the FDTE (e.g., real vs. simulated flight and all types of RPV missions and activities vs. the conduct of only those tasks associated with actual RPV flight missions). The lower levels of workload for the FDTE may also reflect the contributions of the

enhanced software, limited duties, and improved training received by the crew members for the FDTE.

CONCLUSIONS

Two broad conclusions can be drawn from this evaluation of the use of OWL scales under field test conditions.

1. The TLX scale had both the highest factor validity and the best level of operator acceptance within.

2. Operator workload measures may be successfully applied and evaluated.

Both of these conclusions must be viewed relative to the limited number of subjects and the constrained test conditions of the present investigation.

REFERENCES

Bittner, A. C., Jr., & Morrissey, S. J. (1988). Analysis of "error" variances in repeated measures designs. In F. Aghazadeh (Ed.), Trends in ergonomics/human factors: Vol. V. New York, NY: North-Holland.

Dixon, W. J. (Ed.). (1983). BMDP statistical software. Los Angeles, CA: University of California Press.

Hart, S. G., & Staveland, L. E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. S. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Hinkley, D. V. (1983). Jackknife methods. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.) Encyclopedia of statistical sciences: Vol. 4 (pp. 280-287). New York: Wiley.

Reid, G. B., Shingledecker, C. A., & Eggemeier, T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting (pp. 522-525). Santa Monica, CA: Human Factors Society.

Vidulich, M. A., & Tsang, P. S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement application. Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.

DATA ATTACHMENT G-1

AQUILA FDTE II FACTOR SCORES

		MC	AVO	MPO	RPVT	Mean
Crew A						
Mission Planning	Mission 1	-0.58	-0.88	-1.30	-0.53	
	2	-1.38	-1.05	-0.49	-1.25	
	3	-1.30	-1.03	-1.28	-0.64	
	4	-1.20	-1.33	-1.30	-0.28	
						- 0.99
Flight	Mission 1	-0.73	-1.38	0.61	-1.05	
	2	-1.35	-1.43	0.87	-1.43	
	3	-0.35	-1.03	0.35	-0.38	
	4	-1.30	-1.05	1.02	0.17	
						-0.52
	Mean	-1.03	-1.14	-0.19	-0.67	-0.75
Crew B						
Mission Planning	Mission 1	1.12	-1.10	0.95	0.32	
	2	-1.23	-1.15	0.43	0.66	
	3	0.33	-1.10	1.57	0.71	
	4	0.50	-1.20	1.31	-0.71	
						0.09
Flight	Mission 1	1.91	-0.83	1.35	0.19	
	2	1.68	-0.36	1.69	-0.66	
	3	1.79	-0.24	1.46	-0.86	
	4	1.93	-0.42	1.41	-0.93	
						0.48
	Mean	1.00	-0.80	1.27	-0.33	0.28
Crew C						
Mission Planning	Mission 1	1.56	-1.00	0.66	0.31	
	2	0.45	-1.25	0.96	0.06	
	3	0.11	-1.30	-0.18	0.06	
	4	0.11	-1.30	0.67	0.01	
						0.00
Flight	Mission 1	0.48	-1.25	0.88	0.92	
	2	1.81	-1.30	0.37	-1.01	
	3	0.47	-1.30	-0.38	-1.01	
	4	1.42	-1.30	1.09	0.83	
						0.04
	Mean	0.80	-1.25	0.50	0.02	0.02
Crew D						
Mission Planning	Mission 1	-0.02	-0.14	-0.38	-0.71	
	2	0.18	-0.78	0.37	-0.29	
	3	-0.54	0.43	-0.88	-0.71	
	4	0.58	1.26	0.08	-0.64	
						-0.13
Flight	Mission 1	0.45	0.59	0.77	-0.19	
	2	0.40	1.15	0.94	-0.32	
	3	0.38	0.33	0.58	-0.17	
	4	0.65	0.48	1.91	-0.64	
						0.45
	Mean	0.26	0.41	0.42	-0.40	0.17

DATA ATTACHMENT G-2

FACTOR SCORES FOR AQUILA CREW MEMBERS PARTICIPATING IN FDTE II AND OT2

Crew 1

	RPVT		AVO		MC		
	OT2	FDTE2	OT2	FDTE2	OT2	FDTE2	Mean
Mission Planning	1.80	-0.23	-0.47	-1.21	-0.89	-1.40	-0.40
Flight	1.63	-0.30	-0.68	-1.26	-1.40	-1.37	-0.56
Target Detection	1.60	0.10	0.13	-0.58	-0.82	0.33	0.13
Mean	1.68	-0.14	-0.34	-1.02	-1.04	-0.81	-0.28

Crew 2

	RPVT		MPO		
	OT2	FDTE2	OT2	FDTE2	Mean
Mission Planning	1.67	-0.08	1.09	1.05	0.93
Flight	0.65	-1.19	1.96	1.14	0.64
Target Detection	1.54	-0.59	1.40	0.68	0.76
Mean	1.29	-0.62	1.48	0.96	0.78

Crew 3

	MPO		RPVT		
	OT2	FDTE2	OT2	FDTE2	Mean
Mission Planning	0.09	0.49	0.47	0.34	0.35
Flight	0.95	0.36	1.32	0.24	0.72
Target Detection	1.97	0.76	0.55	0.68	0.99
Mean	1.00	0.54	0.78	0.42	.69

Crew 4

	RPVT		AVO		
	OT2	FDTE2	OT2	FDTE2	Mean
Mission Planning	1.35	0.28	-0.35	0.61	0.47
Flight	1.48	0.52	0.98	0.05	0.76
Target Detection	1.43	1.01	1.64	1.10	1.30
Mean	1.42	0.60	0.76	0.59	0.84

APPENDIX H

WORKLOAD ASSESSMENT OF AQUILA REMOTELY PILOTED VEHICLE (RPV) OPERATIONS DURING AN OPERATIONAL EXERCISE *

James C. Byers Richard E. Christ Susan G. Hill
Alien L. Zaklad

ABSTRACT

Operator workload (OWL) assessments were made by operators of the Aquila remotely piloted vehicle (RPV) during a live-fire exercise using two subjective rating scales: Task Load Index (TLX) and Overall Workload (OW). Ratings were made by operators in the ground control station, the remote ground terminal, and the launch and recovery subsystems. Principal components analysis revealed the presence of a single factor – the OWL factor. Analyses of variance examined the effects of several variables on the OWL factor scores and on TLX subscale scores. Significant findings reflect upon the system and its operation. Comparisons are made between these results and OWL assessments made during an earlier Force Development and Experimentation (FDTE) program. These findings are discussed in the context of the development and validation of a methodology for assessing OWL.

INTRODUCTION

This study was designed to evaluate the workload of Aquila remotely piloted vehicle (RPV) operators when the system was used outside of a testing environment and in a situation in which the Aquila was actually being flown. In a previous workload analysis of the RPV during a Force Development Test and Evaluation (FDTE) program (documented by Byers, Bittner, Hill, Zaklad, & Christ, 1988), the RPV was not actually flown but was attached to the underside of a small manned aircraft. (see also Appendix G of this report). Accordingly, the results of this study were compared with those of the previous study.

Background

FIREX 88 was a major live-fire artillery exercise held in June, 1988, at Dugway Proving Ground, Utah. During its employment in FIREX 88, Aquila was employed tactically, for the first time in its history, rather than used in a test and evaluation context. The tactical objectives of the Aquila system during FIREX 88 were to perform target detection, recognition, and location, call for fire, and fire spotting tasks. In addition, an ancillary objective of the Aquila battery was to introduce and demonstrate the capabilities of the RPV system to senior military commanders and other interested parties.

Purpose

The workload study conducted during FIREX 88 was designed to address the following questions.

- What are the relative capabilities of two alternative operator workload (OWL) rating scales when they are administered in the field and in near real time?
- Are the OWL measures obtained sensitive to acknowledged differences in workload resulting from crew positions in the Aquila ground control station (GCS) and mission segments?
- Are the OWL measures obtained sensitive to the workload associated with different components of the Aquila RPV system?
- Are there differences between the OWL data obtained during the FIREX 88 "demonstration" exercise and the Aquila FDTE?

* This appendix contains a revised and condensed version of unpublished Technical Memorandum Number 4, prepared by the indicated authors in 1989.

METHOD

Subjects

The subjects were 15 GCS crew members, three Remote Ground Terminal (RGT) crew members (one also served as a GCS crew member subject), and three launch and recovery subsystem crew members (one also served as a GCS crew member subject). Taking overlaps into account, a total of 19 subjects provided workload ratings.

Each GCS crew consisted of three members: the Mission Commander (MC), the Air Vehicle Operator (AVO), and the Mission Payload Operator (MPO). During FIREX 88, however, there were as many as five crew members working in the GCS, as training in all three duty positions was ongoing. Two chief warrant officers alternated over missions as MC, and the other thirteen GCS crew members (private fist class through sergeant in rank) rotated, somewhat irregularly, as AVOs, MPOs, and trainees for all three crew positions.

The launch and recovery subsystems subjects were two launch and recovery team chiefs and an RPV mechanic. The RGT subjects were an RPV senior non-commissioned officer, an MPO, and an RGT specialist.

Procedures and Instruments

The workload assessment scales used for rating workload were the Task Load Index (TLX) (Hart & Staveland, 1987) and the Overall Workload (OW) scale (Vidulich & Tsang, 1987).

Individual workload ratings were obtained from GCS crew members immediately after the conclusion of each of seven Aquila missions which were conducted over a period of four days. Each of the seven missions had a different crew configuration. Each crew member rated workload using both scales for three or four mission segments. The mission segments were Launch, Flight Operations, Recovery, and when appropriate, the Flight Operation sub-segment of Target Location/Call for Fire.

Individual workload assessments for the RGT and for the launch and recovery subsystems were obtained near the end of FIREX 88. Three individuals rated RGT workload for two mission segments: Power-up and Align. Another three individuals rated launch and recovery subsystem workload for four segments: Activate and Check Out the Launch Subsystem, Conduct Launch, Activate and Check Out the Recovery Subsystem, and Conduct Recovery. The workload assessments for the RGT and the Launch and Recovery subsystems did not reflect workload on any one mission but rather an average workload over all the FIREX 88 missions.

RESULTS

Analyses were conducted in three phases which respectively examined: (a) the factor validities of the two workload scales, (b) the workload associated with different mission segments and RPV components, and (c) the comparison of FIREX 88 workload results with those from the 1987 Aquila FDTE as presented by Byers et al. (1989), and in Appendix G of this report.

Factor Validity Analysis

Principal component analysis (PCA) was conducted on 124 sets of workload ratings across all subjects, systems, and mission segments using BMDP4M (Dixon, 1983). Each set of ratings included global measures of workload using two different scales: TLX and OW. This analysis revealed a single component hereafter called the **OWL factor**, which explained 83.4% of the total variance. This analysis also yielded OWL factor scores which were the basis for the workload analysis reported in the next section. The results of this initial analysis support the view that the two workload scales essentially provide an assessment of a single common factor. (The factor scores for each subject's workload judgments are in Data Attachment H-1 at the end of this appendix).

Workload Analysis

The workload analyses were conducted in three steps corresponding to the three components of the RPV system: the GCS, the RGT, and the launch and recovery subsystems.

GCS workload. Repeated measures analysis of variance (ANOVA) was used to evaluate the effects of Mission Segment (Launch, Flight, and Recovery) and Position (MC, AVO, MPO) on OWL factor scores across all RPV flights. This analysis revealed a significant segment-by-position interaction ($F(4,52) = 5.48$, $p < 0.001$. This interaction is illustrated in Figure H-1. It may be

noted that while the MC has the highest and a relatively constant OWL factor score across mission segments, the workload ratings of the AVO and MPO vary inversely from each other from segment to segment.

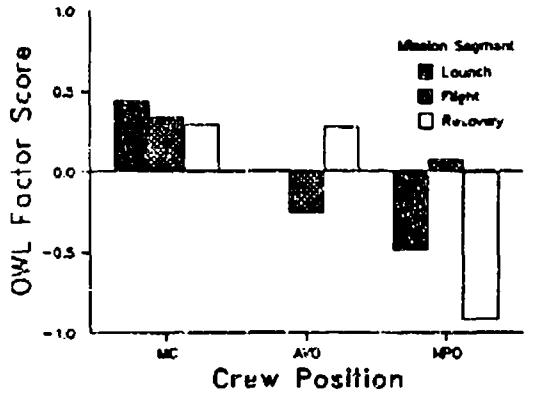


Figure H-1. The effect of mission segment and crew member positionon workload.

An ANOVA of the same structure but using TLX subscale ratings in place of OWL factor scores also reveals the segment-by-position interaction ($F(4,52) = 4.15$, $p < .01$), as well as significant effects for subscale ($F(5,130) = 16.52$, $p < 0.0001$) and the segment-by-position-by-subscale interaction ($F(20,260) = 2.70$, $p < 0.0005$). The subscale main effect is caused by variations in mean weighted subscale ratings: the mean rating for Mental Demand (190) was the highest, followed by those for Temporal Demand (141), Frustration (129), Performance (99), Effort (84), and Physical Demand (14). [Note that the weighted subscale scores can range from 0 to 500 depending on the subscale rating value (0 to 100) and the magnitude of the subscale weight (0 to 5).]

The three-way interaction showed that the subscale ratings varied as a function of the joint effect of variations in crew member position and mission segments. While there are several possible instances of these joint effects, one of the more obvious is the relatively high levels of Mental and Temporal Demand reported by the MC in all three mission segments, and the shifts in these two components of workload for the AVO and MPO as a function of mission segments. In particular, the MPO reported higher levels of Mental and Temporal Demand than the AVO for Flight segments, while the AVO had higher levels of these two workload components than the MPO during Launch and Recovery operations. These results mirror the Mission Segment-by-Crew member position interaction effects on OWL factor scores shown in Figure H-1.

RGT workload. An ANOVA examined the effect of two RGT mission segments, Power Up and Align, on OWL factor scores across three RGT crew members. No significant effects were found. Another ANOVA checked the effects of the two RGT mission segments on TLX weighted subscale scores. Only the subscale main effect was found to be significant, ($F(5, 10) = 6.60$, $p < 0.01$). The highest subscale score was for Temporal Demand (397), followed in order by Performance (161), Physical Demand (158), Effort (149), Mental Demand (61), and Frustration (27).

Launch and recovery subsystem workload. An ANOVA was used to evaluate the effects on OWL factor scores of two types of tasks (Activate and Check out a subsystem and Conduct AV operations using the subsystem) and two types of subsystems (Launch and Recovery). A significant effect was found for Task ($F(1,2) = 78.18$, $p < 0.02$). Mean OWL factor scores were higher for the task of Activating and Checking out a subsystem (.48) than for the task of Conducting Operations with the subsystem (-.42). The mean OWL factor score for the Launch subsystem (0.35) was higher than that for the Recovery subsystem (-0.27), but the subsystem main effect was not significant ($F(1,2) = 7.9$, $p > .10$).

An ANOVA conducted to assess the effects of two types of tasks and two subsystems on TLX weighted subscale scores revealed a significant effect for Subscale ($F(5,10 = 3.63$, $p < 0.04$). As was the case for the RGT data, the highest mean subscale score for the launch and recovery subsystems was for Temporal Demand (273). However, the ordering of the other subscales by their respective values was different. For the launch and recovery subsystems the order of subscales after the Temporal Demand was Frustration (152), Effort (89), Physical Demand (88), Mental Demand (72), and Performance (49). The mean weighted TLX subscale scores of the launch and recovery subsystems (139 and 102, respectively), track the differences found for the OWL factor scores.

Comparison of Workload During FIREX 88 and the FDTE

An ANOVA was used as the basis for

comparing OWL factor scores from the present FIREX 88 study with those reported from the Aquila FDTE by Byers et al. (1988). The analysis was limited to the subjects who served as crew members (in any crew position) in both studies and to workload ratings for the GCS mission segment of Flight Operations, which was the only rating common to both studies. This analysis revealed a significant test-by-position interaction ($F(2, 92) = 3.03$, $p < 0.05$), as illustrated in Figure H-2. It may be seen in the figure that for the AVO, the mean OWL factor score is higher for FIREX 88 than for the FDTE (though below the average OWL factor score in both cases). For the MC and MPO, the opposite is true.

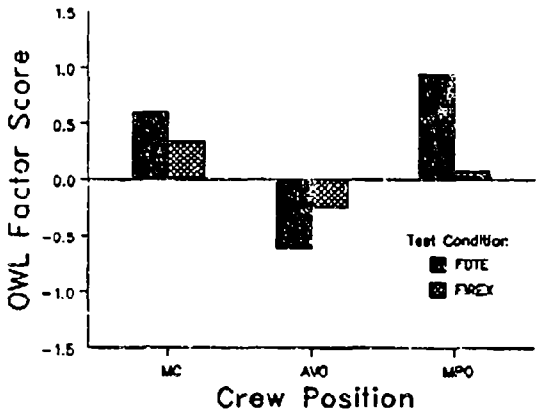


Figure H-2. The effect of test condition and crew member position on workload.

DISCUSSION

The TLX and OW workload assessment scales were successfully applied in investigation of the workload experienced by operators of the Aquila RPV during FIREX 88. The nature of the Aquila role at FIREX 88 was many sided. The RPV flights were used for providing specific types of support to the field artillery, for training system operators for new duty positions, and for providing general publicity on the capabilities of the Aquila system to any and all interested individuals and agencies. Despite the presence of trainees and many visitors in the GCS and around the other subsystems, and many last minute changes in flight purposes and plans, the application of the two scales revealed a coherent picture of operator workload in three Aquila subsystems.

GCS Workload Evaluation

The workload analyses indicated significant mission segment-by-crew position interaction. The nature of the interaction is entirely consistent with the nature of the roles of the crew members during a mission. The MC has a fairly constant high level of workload which probably reflects the constant high level of responsibility over the entire mission. The AVO has the least workload in fight segments during which his assigned tasks are fairly routine, and the greatest workload in the recovery segment during which great pressure is placed on the AVO to "put the bird into the net," a task requiring the preparation and execution of a precise and time-dependent flight profile. It was not unusual for several factors to arise during this critical maneuver which were capable of sabotaging a successful recovery. The MPO has low workload in launch and recovery segments of a mission (where the mission payload is not in use) and higher workload in the flight segment when the payload is used to detect, recognize, locate, and designate targets.

The TLX subscale main effect was significant, with Mental Demand having the highest mean value as might be expected given the nature of GCS operations. The mean high score on the Frustration subscale is consonant with the FIREX conditions, including the trainees in many positions, visitors walking into and out of the GCS, and various problems with communications. The segment-by-position interaction for TLX subscale values shows clear differences in the sources of workload across mission segments and crew member positions.

RGT and Launch/Recovery Subsystem Workload Evaluation

Though a limited sample size restricts the usefulness of the analyses of workload associated with operating the RGT and the launch and recovery subsystems, several interesting results are apparent. First of all, while Mental Demand is the largest component of workload in the GCS, the main driver of workload in the RGT and the Launch and Recovery subsystems is Temporal Demand. The high Frustration level for the launch and recovery team was, as observed by the assessment team, mainly due to the difficulty incurred in trying to maintain and operate first generation, prototype equipment. Secondly, the workload of the launch/recovery team is higher for

the activate and check out task than in the actual conduct of launch and recovery. This finding again reflects the problem inherent in working with prototype equipment; once it is "up and running" it is not difficult to operate, but it is often difficult to get it to that desirable state. Finally, the data support the contention that launch operations involve more workload than does recovery.

Comparison of FIREX and FDTE Workload

Comparing workload measures obtained for flight operations segments of Aquila missions during FIREX and the FDTE, the expectations would be for the AVO to have higher workload in FIREX (because the RPV is actually being flown, for the MPO to have lower workload in FIREX (because target detection was not a major objective of the flights), and for the MC to have lower workload in FIREX (because the MCs in FIREX were much more experienced than the MCs in the FDTE and because the pressure to perform flawlessly during FIREX was not as great). The comparison of FIREX and FDTE workload assessments confirms these expectations.

CONCLUSION

Several conclusions may be drawn from the present evaluation of the use of the OWL scales under field conditions.

1. OWL measures may be successfully applied within the field exercise environment as found at FIREX 88.

2. OWL on the RGT and the launch and recovery subsystems is principally due to time pressure.

3. OWL in the GCS varies by mission segment and crew member position.

4. The AVO has more workload when the RPV flight is actual rather than simulated.

REFERENCES

Byers, J. C., Bittner, A.C., Jr., Hill, S. G., Zaklad, A. L., & Christ, R.E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1145-1149). Santa Monica, CA: Human Factors Society.

Dixon, W. J. (Ed.). (1983). BMDP statistical software. Los Angeles, CA: University of California Press.

Hart, S. G., & Staveland, L. E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. S. Hancock, & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Vidulich, M. A., & Tsang, P. S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

DATA ATTACHMENT H-1

AQUILA FIREX 88 FACTOR SCORES

Mission 1

	MC	AVO	Trn AVO	MPO	Trn MPO	Mean
Launch	1.01	1.00	1.51	0.04	-2.00	0.31
Flight	0.70	0.52	1.43	0.44	-0.50	0.52
Recovery	1.01	1.17	1.07	0.52	-2.00	0.35
Mean	0.91	0.90	1.34	0.33	-1.50	0.39

Mission 2

	MC	Trn MC	AVO1	AVO2	Trn MPO	Mean
Launch	0.97	0.78	0.01	0.26	-0.49	0.31
Flight	1.19	-0.27	-0.22	0.32	-0.38	0.13
Recovery	1.17	0.17	1.10	0.98	-1.00	0.48
Mean	1.11	0.23	0.30	0.52	-0.62	0.31

Mission 3

	MC	AVO	MPO	Trn MPO	Mean
Launch	0.45	0.77	0.16	-1.00	0.10
Flight	1.36	0.39	1.13	-2.00	0.22
Target D	1.40	-0.82	--	-2.00	-0.36
Recovery	0.51	-1.00	-2.00	-0.26	-0.69
Mean	0.93	-0.17	-0.18	-1.32	-0.18

Mission 4

	MC	Trn MC	Trn MC	AVO	MPO	Mean
Launch	0.46	-0.04	1.38	0.15	0.34	0.46
Flight	0.58	0.00	1.42	-0.08	1.21	0.63
Target D	-0.46	0.05	1.59	0.79	1.87	0.77
Recovery	0.98	0.43	0.84	0.54	-0.99	0.36
Mean	0.39	0.11	1.31	0.35	0.61	0.55

Mission 5

	MC	AVO	MPO	Mean
Launch	-0.56	-2.00	-0.76	-1.11
Flight	-0.78	-2.00	-0.37	-1.05
Target D	-1.00	-2.00	0.39	-0.87
Recovery	-0.95	-1.00	-1.00	-0.98
Mean	-0.82	-1.75	-0.44	-1.00

DATA ATTACHMENT H-1 (Continued)

Mission 6

	MC	AVO	MPO	Trn MPO	Mean
Launch	0.14	0.57	0.93	-0.22	0.36
Flight	0.12	0.12	1.12	0.39	0.44
Target D	0.18	-0.11	-0.15	0.11	0.01
Recovery	-0.56	0.18	1.06	-2.00	-0.33
Mean	-0.03	0.19	0.74	-0.43	0.12

Mission 7

	MC	AV/MPO	MP/AVO	Mean
Launch	-0.17	-2.00	-2.00	-1.39
Flight	-0.90	-2.00	-0.49	-1.13
Target D	-0.55	-2.00	-2.00	-1.52
Recovery	-0.59	-1.00	-0.02	-0.54
Mean	-0.55	-1.75	-1.13	-1.14

Mission 1-7

	L/R Tm Chf 1	L/R Tm Chf 2	RPV Mech	Mean
Activate Launch	1.22	0.47	0.73	0.81
Conduct Launch	0.38	-0.49	-0.23	-0.11
Activate Recovery	0.04	0.04	0.45	0.18
Conduct Recovery	-0.57	-0.86	-0.74	-0.72
Mean	0.21	-0.17	0.04	0.04

Mission 1-7

	Tm Ldr	RGT Crew	MPO	Mean
Power Up RGT	1.26	0.32	0.99	0.86
Align RGT	1.83	0.72	-0.05	0.83
Mean	1.55	0.52	0.47	0.85

APPENDIX I

OPERATOR WORKLOAD ASSESSMENT OF THE UH-60A BLACK HAWK SYSTEM

Helene P. Iavecchia Paul M. Linton Regina M. Harris
Allen L. Zaklad James C. Byers

Abstract

An empirical study was undertaken to collect workload ratings of pilots and copilots performing a resupply mission in a UH-60A flight simulator. Real-time overall and peak workload (OW and PW) ratings were collected for twelve segments of essentially identical day and night missions. Real-time ratings for day missions were compared with OW and PW values predicted by the Task Analysis/Workload (TAWL) and TAWL Operating System Simulation (TOSS) model. Additional post-mission workload ratings using OW, PW, Task Load Index (TLX), Subjective Workload Assessment Technique (SWAT), and Modified Cooper-Harper (MCH) techniques, along with other subject inputs, were also collected. The TAWL/TOSS-derived estimates of workload were highly correlated with real-time workload ratings. Jackknife factor analysis of the post-mission workload ratings revealed the presence of only a single factor (accounting for over 71% of the variance). These and other findings of this study are discussed in the context of the development and validation of a methodology for assessing workload.

INTRODUCTION

The ability to predict and evaluate operator workload (OWL) has become a serious concern as military systems become increasing complex. The OWL Program was an exploratory development program sponsored by the U.S. Army Research Institute (ARI) for the application and validation of practical methods for assessing OWL in Army systems throughout their life cycle. Following study plans documented by Bittner et al., 1987, workload data were collected for three Army systems in varying stages of development. These systems were the Aquila Remotely Piloted Vehicle, the Line-of-Sight-Forward-Heavy (LOS-F-H) component of the Forward Area Air Defense System (FAADS), and the system of interest in this report, the UH-60A BLACK HAWK helicopter.

This report summarizes and documents the OWL Program studies conducted in an Army aviation setting. The primary intent of this effort was to examine the relationship between workload predicted by an analytical model and workload reported by crew members in an "operational setting." Additionally, this study sought to continue the OWL Program investigations into alternative workload rating techniques and analyses of workload associated with Army systems. In performing these studies, the "ideal" operational setting would have been an actual aircraft with the crew flying well-defined, pre-briefed missions. However, the scope of this project precluded the time and expense associated with dedicated flight testing. In lieu of an actual flight test, an Army training simulator was made available for the study, specifically the UH-60A 2B38 flight simulator located at the U.S. Army Aviation Center, Ft. Rucker, Alabama.

Purpose

The objectives of the UH-60A workload studies were to:

This appendix contains a revised and condensed version of unpublished Technical Memorandum Report 2075-4c, prepared by the indicated authors in December, 1989. A paper based on part of this report was presented at and is published in the Proceedings of (pp. 1481-1485) the 33rd Annual Meeting of the Human Factors Society.

- Determine the relationship between an analytical model's prediction of workload and the workload reported by the pilot and copilot while flying a simulated daylight mission,

- Investigate various methodological issues in assessing workload including differences in workload reported during the mission versus workload recalled following the mission, factor validity of the workload measurements, diagnostic capabilities of the data, and operator acceptance of the various assessment techniques, and

- Evaluate the effects of key mission variables on pilot and copilot workload as well as the relationship between performance and workload.

UH-60A System Description

The U.S. Army's UH-60A Black Hawk is a twin-engine rotary-wing utility helicopter designed specifically for combat and combat support missions comprised of tactical transport of soldiers, troop units, and required supplies and equipment. Cockpit, instrument panels, and interior lighting are all designed to accommodate both day and night full-mission capability. The flight control system provides maneuverability for low level, nap-of-the-earth flying. The basic UH-60A crew consists of a pilot, copilot, and crew chief/gunner. The aircraft has virtually identical control and display configurations on either side of the tandem cockpit, and can be properly flown by either the pilot or copilot.

The UH-60A 2B38 flight simulator consists of a molded two-piece cockpit mounted upon a large motion platform. The front cockpit is a faithful reproduction of the fielded UH-60A unit consisting of a pilot and copilot station; behind the flight stations is an instructor/operator station, and an observer station. The cockpit assembly is mounted upon a motion system which provides dynamic movement and accurate cues for pitch, roll, and yaw, along the vertical, lateral, and longitudinal axes, as well as any combination thereof. Four out-the-window cathode ray tube displays are provided for the pilot and copilot stations. The displays allow forward and side viewing of a simulated environment during dawn, day, dusk, night, and night vision goggle (NVG) conditions.

METHOD

OWL Measures

Empirical measures of OWL. Five operator workload rating scales were used: the four workload rating scales selected for evaluation in all of the OWL Program studies. These ratings scales were: (a) Task Load Index (TLX), Hart and Staveland, 1987; (b) Subjective Workload Assessment Technique (SWAT), Reid, Shingledecker and Eggemeier, 1981; (c) Modified Cooper-Harper (MCH), Wierwille and Casali, 1983; (d) Overall workload (OW), Vidulich and Tsang, 1987; and (e) a scale developed specifically for this study, Peak Workload (PW), modelled after the OW scale.

The TLX is composed of six components, each of which contributes to workload. The TLX components -- mental demand, physical demand, temporal demand, performance, effort, frustration -- are also individually rated on a 100-point scale. SWAT measures three workload components -- time, effort, and stress -- with each measured on a three-point scale. Both TLX and SWAT require additional data collection on individual subjects prior to the experimental procedures. MCH uses a decision tree structure to direct the subject to the appropriate workload rating using a ten-point scale. OW is a rating of the subject's overall workload experienced during a particular segment on a unidimensional scale of 0 to 100 with 0 representing very low and 100 representing very high workload. PW is a measure of the "peak workload" experienced during a segment on a scale of 0 to 100. The PW measurement scale was constructed for this study to tap momentary overloads. The concept of peak workload is important in that even one instance of momentary overload can lead to mission failure in certain situations, especially in an aviation setting.

Analytical measures of OWL. The analytical model chosen to make predictions of workload was based on the TAWL/TOSS technique (Bierbaum, Fulton, & Hamilton, 1989). This model was selected for use in this study because its previous applications included the UH-60A (Bierbaum, Szabo, & Aldrich, 1987). This analytical tool requires inputs which include: (a) a detailed task analysis defining the low-level task activities required for each mission-essential task (e.g., control altitude or perform cockpit communication) together with the task times; (b)

estimates of the level of workload in each of five information processing channels (i.e, auditory, visual, kinesthetic, cognitive, and psychomotor) for each low-level task on a scale from 0 to 7 (very low to very high workload); and (c) a set of scenario decision rules to drive the tasks to be performed during each half-second simulation time interval, to include the probability of random concurrent tasks. Given these inputs and the generated time line of low-level task activities, TAWL/TOSS sums the workload values within each channel across concurrent tasks. If the sum of channel workload values (e.g., visual) within a half-second interval exceeds a value of 7, an overload is defined to have occurred for that channel during that interval.

Simulator Data Collection Effort

One week prior to the simulator data collection effort the crew members met as a group for a four-hour prebrief. During this prebrief subjects were told of the intent of the study, given an introduction to the concept of workload and a description of the specific methods that would be used in the current study to measure workload. A questionnaire was also administered to the subjects during the prebrief period to gather information concerning the subjects' experiences in flying. The questionnaire also provided the aviators with an opportunity to use the OW and PW rating scales by recalling and rating their past experiences during particular missions (day or night) and mission segments. Finally, pretest data necessary to use the two multidimensional scales (i.e., the TLX and SWAT scales) were collected at this time.

The data collection test conditions are summarized below:

- Real-time verbal reports of OW and PW by the pilots and copilots during the simulator flight,
- Real-time performance assessment of the crew by an instructor pilot observing the simulator flight, and
- Post-time ratings of workload by the pilots and copilots during a mission debrief including the OW, PW, SWAT, TLX, and MCH scales.

Subjects. Ten two-man crews participated in the study. All subjects were experienced UH-60A aviators and were currently assigned as instructor pilots (IPs) at the U.S. Army Aviation Center. Two additional senior IPs were selected to rate the performance of the pilot and copilot during the simulator trials and to assist in the collection of real-time pilot and copilot workload ratings.

UH-60 missions. Each crew flew two experimental flights -- one day mission and one night vision goggle (NVG) mission. Half the crews flew the day mission first and half the NVG mission first. The two missions were essentially the same although the night mission was confined to a smaller, as well as different, geographical area to accommodate the slower speeds flown at night. In both flights, the crew flew a one-hour resupply mission in the UH-60 flight simulator. The mission required a team of two BLACK HAWKS to navigate to a pick-up point, hook up an external sling load of fuel blivets, and deliver the cargo to a forward drop-off point. At the start point, the experimental crew was notified that the second BLACK HAWK experienced an equipment malfunction and they were to complete the mission in a stand-alone role. This necessitated an alternate drop-off point, and an unanticipated visit to a forward arming and refueling point (FARP). Threats were simulated at selected mission segments (4, 6, 8, and 10) along with an engine out emergency. The mission segments and their abbreviated codes are listed in Table I-1.

Crew procedures. During the simulated experimental flights, the primary task of the pilot was limited to flight management and that of the copilot to navigation and communications. Once a mission was underway, the controller IP asked both operators to report in near real-time the OW and PW experienced during each of twelve mission segments. The controller IP also rated the performance of both operators for each segment. The scale used for rating performance is similar to the one normally used by IPs while evaluating candidate aviators during training. Following each experimental flight, the two crew members gave retrospective workload ratings for all twelve mission segments using the OW and PW scales and for only four selected mission segments (Segments 3 through 6) using the TLX, SWAT, and MCH techniques. Following the post-mission period of rating workload, a structured interview was conducted with both crew members to assess operator acceptance of the various rating techniques and to gather other general comments.

TAWL/TOSS Data Collection Effort

The baseline UH-60A model (Bierbaum et al., 1987) was updated to include all the pilot and copilot task activities that were employed by the crews during the experimental flights which occurred during daylight. The decision rules that control when the pilot and copilot tasks are triggered during the TAWL/TOSS simulation were also updated to reflect the specific mission requirements of the experimental flight. This updating effort was independently accomplished by Anacapa Sciences, Inc. (D. B. Hamilton and C. R. Bierbaum, personal communication, December, 1989). Following the updates, a copy of the UH-60 application code as well as the TAWL/TOSS software Version 2.0 were delivered to the authors of this report for execution.

Because TAWL/TOSS is stochastically based, it was necessary to run the model a number of times and average the results. For this study, the model for daylight operations was executed seven times and the average output of the runs was used in a comparison with the crew data collected in the experimental daylight flights. Since TAWL/TOSS does not directly generate OW and PW values, it was necessary to develop a procedure to derive these values. To derive a TAWL/TOSS-based estimate of OW for each mission segment, the TAWL/TOSS workload values for each half-second interval within a mission segment were averaged over all five TAWL/TOSS channels (i.e., auditory, visual, etc.). The derived (or predicted) OW score was the mean of these half-second values over the duration of the mission segment. To derive a TAWL/TOSS-based estimate of PW for each mission segment, the TAWL/TOSS workload values for each half-second interval were summed across the five TAWL/TOSS channels. The maximum value of all half-second summed values was defined as the PW for that segment. All TAWL/TOSS derived OW and PW scores were scaled to correspond with the 0 to 100 scale used by the crews to rate workload in the simulated experimental flights.

RESULTS

The results are presented in three major sections in accordance with the goals of the study: (a) TAWL/TOSS predictions of crew workload, (b) methodological issues in workload assessment, and (c) UH-60A workload issues. With the exception of the operator questionnaire, three crews were eliminated from the analysis of results. One crew did not complete the study due to extreme simulator sickness experienced by one of the crew members. Two other crews were excluded because the crew members altered pilot and copilot responsibilities, thereby creating workload conditions that differed from the other crews who flew with well-defined and fixed pilot and copilot roles.

TAWL/TOSS Predictions and Operator Ratings of Workload

Results for six of the twelve mission segments were analyzed (Segments 3, 4, 5, 8, 11, and 12). Other segments were not considered due to missing data (Segments 6 and 10), simulator failures (Segments 1 and 2), and repetitive types of segments (Segments 3 and 7 are both Pickup Zone (PZ) operations, Segments 5 and 9 are both Landing Zone (LZ) operations). The average ratings of the pilots and copilots and the TAWL-derived values for each applicable segment is in Data Attachment I-1 at the end of this appendix.

Figure I-1 graphically illustrates the comparison of average OW ratings with the TAWL\TOSS predicted OW scores as a function of mission segment, separately for the pilot and copilot. The correlation across all crew members between real-time ratings and predicted OW scores was significant ($r = 0.82$; $p < .01$). As shown in Figure I-1, TAWL/TOSS predictions track the OW ratings across segments. However, with one exception, the real-time OW ratings are higher than the TAWL/TOSS-based workload prediction ($F(1,10) = 6.81$, $p = 0.026$). The exception is the pilot's OW rating for PZ Operations -- Segment 3 For this case, the TAWL/TOSS model predicted higher workload than reported. This may be due to the fact that the pilot communication was not as complex as was originally assumed in the TAWL UH-60 model (D. B. Hamilton and C. R. Bierbaum, personal communication, January, 1990). It is noteworthy that the correlation between TAWL and the real-time OW ratings increases from 0.82 to 0.95 without the pilots' data for this segment.

While statistically significant, the TAWL/TOSS-derived PW scores did not predict the crew reports as well as the TAWL/TOSS-derived OW predictions ($r = 0.62$; $p < .05$). The PW predictions are flatter than the real-time PW ratings of both the pilot and copilot. That is, the predicted PW values frequently do not discriminate

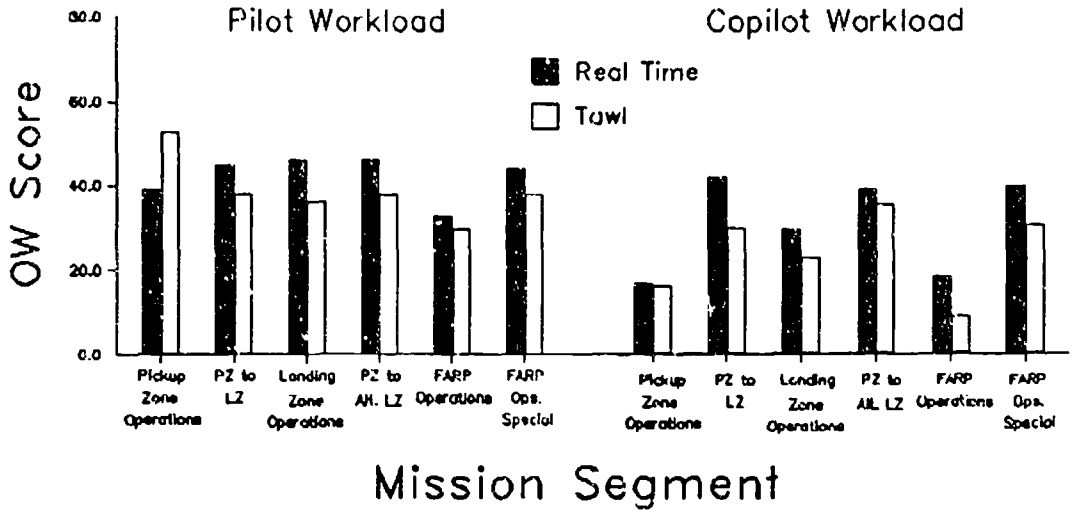


Figure I-1. The real-time ratings and the TAWL/TOSS model predictions of UH-60A global workload as a function of mission segment and crew member position.

differences in workload between segments as reported by the pilots and copilots. Indeed, four of the six TAWL PW predictions are identical in both the pilot and copilot cases. Furthermore, in contrast to OW, the TAWL-derived measures of PW also overestimated the PW reported by the crews.

Methodological Issues in Workload Assessment

OW and PW Scales. An analysis of variance (ANOVA) was conducted to determine the effect on workload ratings of the two rating scales (OW and PW), two rating times (real-time verbal reports and post- mission written reports), two missions (day and night), ten segments (1, 2, 3, 4, 5, 7, 8, 9, 11 , and 12), and two crew position (pilot and copilot). (The mean ratings for combinations of these conditions data are given in Data Attachment I-1.) The main effects of all of these factors except crew position were significant.

The mean OW score was 39.1 and the mean PW score was 48.0 ($F(1,12) = 82.4$, $p < .0001$). The average real-time rating (46.0) was higher than the average post-mission rating (41.0), ($F(1,12) = 5.97$, $p < .03$). The average workload rating for day mission (37.3) was lower than that for NVG missions (49.8) ($F(1,12) = 29.33$, $p < .0002$).

The mean ratings for each of the segments are shown in Table I-1 ($F(9,10) = 15.7$, $p < .0001$). The greatest workload was found in Segment 12, the segment in which an engine failure occurred enroute from the FARP to the start point. The segments in which the crew flew between the pickup zone and the landing zone with the external fuel bliver load (Segments 4 and 8) were also rated as high in workload relative to other segments. Refueling at the FARP (Segment 11) as well as the two initial flight segments (Segments 1 and 2) enroute to the pickup zone had lower workload ratings.

The ANOVA of OW and PW ratings also revealed several significant interactions. The Scale-by-Segment ($F(9,108) = 12.55$, $p < .0001$), Segment-by-Position ($F(9,108) = 5.40$, $p < .0001$), and Scale-by-Segment-by-Position ($F(9,108) = 3.96$, $p < .0002$) interactions indicate that workload ratings varied as a function of varying combinations of the Rating Scale, Mission Segment, and Crew Position. The difference between the two scales, always showing PW greater OW, was fairly constant in magnitude except for Segment 12 which included the simulated engine failure; the PW ratings were particularly greater than the OW ratings for this segment. The average workload ratings of pilots were always at least moderately greater than those for copilots but were substantially so on five of the 10 mission segments analyzed: both PZ Ops Segments (3 and 7), the LZ Ops and alternate LZ Ops (5 and 9, respectively), and the FARP Ops (11). An explanation of the three-way interaction among these factors is not clear but are due in part to a much greater difference between OW and PW ratings for the copilot in Segment 12 than for the pilot.

Table I-1

Mean Real-time Workload Ratings for Mission Segments in the UH-60A Simulation Study

Segment Number	Description	Code	Rating
1	Startpoint to Checkpoint 1	SP-CP1	36.0
2	Checkpoint 1 to Pickup Zone	CP1-PZ	38.4
3	Pickup Zone Operations	PZ Ops	42.5
4	Pickup Zone to Landing Zone	PZ-LZ	50.4
5	Landing Zone Operations	LZ Ops	46.3
6	Landing Zone to Pickup Zone	LZ-PZ	**
7	Pickup Zone Operations	PZ Ops	40.9
8	Pickup Zone to Alternate LZ	PZ-Alt LZ	49.5
9	Alternate LZ Operations	Alt LZ Ops	48.6
10	LZ to Forward Arming & Refueling Point (FARP)	LZ-FARP	**
11	FARP Operations	FARP Ops	31.5
12	FARP to Special Including Engine Failure	FARP-SP	52.9

Note. Segments 6 and 10 are not included due to missing data.

The only other significant effect for OW and PW ratings was a Segments-by-Rating Time-by-Mission interaction ($F(9,108) = 1.98$, $p < .05$). This interaction may be attributed to a greater difference between real-time and post-mission ratings for NVG missions than for day missions.

Factor validity of alternate rating scales. Principal component analysis (PCA) was conducted on 160 sets of workload ratings using BMDP4M (Dixon, 1983). Each set contained the ratings obtained using four scales: TLX, OW, MCH, and SWAT. For comparative purposes, these four scales were chosen to match those used in the other Army system studies conducted for the OWL Program. The analysis revealed a single component, hereafter called the OWL factor, which explained 71.4% of the variance. This result indicates that all four workload scales provide assessments of what is essentially a single common factor. Jackknife PCAs were conducted to evaluate the stability of the factor loading of the four workload scales (i.e., correlations with the OWL factor). Jackknife analysis involves successively dropping subjects, one-at-a-time, from a data set to examine the stability of parameter estimates (Hinkley, 1983). An ANOVA of the jackknife results revealed a significant difference among the scale factor loadings ($F(3,57) = 1165.8$, $p < .0001$). Subsequent analysis revealed the following ordering of the factor loadings:

TLX(.899), OW(.872), SWAT(.805), MCH(.795).

All differences are significant with the exception of the difference between SWAT and MCH.

Analysis of TLX subscale results. An ANOVA was conducted to determine the effects on workload ratings of the six TLX subscales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration), four mission segments (3, 4, 5, and 6), two missions (day and NVG), and two crew positions (Pilot and Copilot). The analysis was conducted using the TLX weighted subscale scores. (These data are given in Data Attachment I-2.)

The main effect for each of these four factors was shown to be significant. The ordering of weighted TLX subscale values was Mental Demand (115), Temporal Demand (112), Effort (109), Performance (62), Physical Demand (40), and Frustration (32), $F(5,60) = 9.19$, $p < .0001$. Clearly, the major contributors to global workload ratings were due to the first three of these subscale values.

The other three main effects have previously been examined in terms of their effect on OW and PW ratings. For two factors, the results here show that TLX ratings are affected in about the same way as OW and PW ratings. The average weighted TLX subscale scores for Segments 3 through Segment 6 was 73, 93, 77, 70, respectively, $F(3,36) = 3.88$, $p < .02$. The workload associated with Segment 4 (enroute from pickup zone to landing zone with the external fuel blivet load) was

greater than the other three segments. The mean TLX subscale value for day missions (70) was lower than that for NVG missions (86), $F(1,12) = 9.99$, $p < .01$. These TLX subscale data revealed, in contrast to the OW and PW ratings, that Pilot workload (98) was significantly higher than Copilot workload (58), $F(1, 12) = 5.63$, $p < .05$.

Two interactions were also revealed to be significant. The interaction between mission and mission segment, $F(3,36) = 4.01$, $p < .05$, is due to the fact that workload is significantly lower for Day Missions than NVG Missions except for Segment 4 where there is no difference. The mission segment-by-TLX subscale interaction, $F(15,180) = 2.51$, $p < .002$, is illustrated in Figure I-2. This figure illustrates the generally higher workload in Segment 4 than Segments 3, 5, and 6, but furthermore

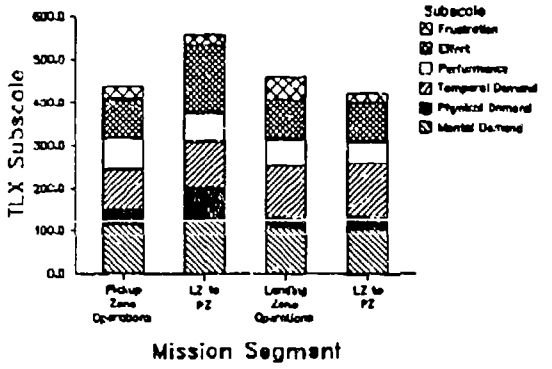


Figure I-2. The effect of mission segments and TLX subscales on workload scores in the UH-60A study.

indicates that the source of the higher workload is principally due to increases in Physical Demand and Effort. This result is reasonable considering that the crew is flying through hostile territory and that the platform can become unstable while carrying the heavy external load. The high level of Physical Demand can be attributed to vibrations in the platform that interfere with fine motor control and/or to physiological responses to stress.

Performance and workload. An analysis was conducted to examine the relationship between the crews' real-time workload measures (OW and PW) and the independent rating of performance (IRP) made by the senior IP who observed the missions. No significant relationships were found between workload ratings and the IRP ($r = 0.0$ for the correlation of IPR to OW and PW).

Operator acceptance of workload assessment techniques. Results of a questionnaire concerning crew acceptance of the five workload assessment techniques employed in this study were analyzed. The pilots were asked four questions about scale usage. These questions and the results are presented in Table I-2. For every question, the pilots rated each workload rating technique on a five point scale. For Questions 1 through 4, respectively, a rating of 1 represented the most favored technique, the easiest technique, the most difficult technique, and the best technique for describing workload experiences. The data presented in Table I-2 are the mean rating response of the crew members. The OW scale was liked the best and was also rated easiest to use. The MCH scale was rated the hardest to use. Finally, TLX was rated highest as the scale that best allowed the crew members to rate the workload they experienced. An interesting comment on the use of the PW scale was that it required more time to respond to because all the events in a segment had to be recalled before a PW value could be determined.

TABLE I-2

Operator Acceptance of Workload Rating Scales in the LOS-F-H NDICE Study

	Rating Scale			
TLX	OW	PW	MCH	SWAT
Which of the questionnaires did you like the best?				
2.7	1.9	2.5	4.1	3.6
Which questionnaire was the easiest to fill out?				
3.2	1.7	1.9	4.0	3.8
Which questionnaire was the hardest to fill out?				
2.5	3.8	3.8	2.1	2.4
Which questionnaire do you think best allowed you to describe the workload you experienced?				
2.2	2.6	2.8	3.8	3.1

Note. Data shown are the mean rating for each scale, where 1 is the most favorable rating.

UH-60 Crew Member Workload

This section focuses on the segments with the highest reported workload (real-time OW and PW ratings) for the pilot and copilot. A TLX evaluation is also provided for those segments where TLX data were available (Segments 3, 4, 5, and 6 only.)

Pilot workload. For the day mission, the segments with the highest real-time pilot OW were, in order of highest to lowest, Segments 9, 8, and 5 (Alt LZ Ops, PZ-Alt LZ, and LZ Ops, respectively. These results are in line with the pilots' comments collected during the post-mission debriefs. Specifically, the pilots noted that LZ and PZ operations had the greater workload. There are several reasons why the PZ-Alt LZ segment had high workload. First, at the start of this segment, the crew was notified of a mission change -- the blivets were to be taken to an alterative landing zone. This required immediate navigation planning. Second, it is to be expected that high workload be associated with carrying the external fuel blivets through hostile territory. To avoid enemy detection, the pilot must fly close to the ground while the blivets are suspended below the helicopter on a cable. An explosion could result if the blivets collide with the ground. Also, as previously mentioned, the platform can become unstable if excessive oscillation of the heavy load exceeds the control system's ability to maintain stable flight.

The highest real-time PW ratings for the day mission were in line with the OW ratings with one exception: Segment 12 (FARP-SP) moved into a second place ranking for PW ratings (Segments 8 and 5 shifted to fourth and fifth place). Relatively high momentary workload would be expected for FARP-SP because of the engine failure which occurred during this last segment of the mission.

For the night mission, the highest OW was experienced in Segments 5, 3, and 7 (LZ Ops, the first PZ Ops, and the second PZ Ops, respectively). The PZ and LZ Ops were more difficult at night because of the reduced visibility. There was a much greater danger of collision with trees or other objects in the landing areas. With the same exception as was true for day missions, the real-time PW ratings at night were in line with the OW ratings at night. Again, the one exception was for Segment 12 (FARP-SP); this segment which included the emergency situation was given a relatively high real-time PW rating.

The TLX results available for Segments 3 through 6 provide some information concerning factors which contribute to workload ratings. The TLX subscale results revealed that, for the pilot, of the three highest rated components, Mental Demand and Temporal Demand were greater than Effort in their contribution to overall workload (152 and 115, respectively). This difference was, if anything, greater for night missions than for day missions. At night, the greater impact of Mental and Temporal Demands are even more pronounced than they are during the day. This latter observation is probably due to the fact that there is less visibility at night and therefore less time and more mental demand to avoid collisions with landing zone objects.

Copilot workload. For the day mission, the three segments in which the copilots experienced the highest real-time OW were Segments 4, 12, and 8 (PZ-LZ, FARP-SP, and PZ-Alt LZ, respectively). The highest copilot real-time PW during daylight missions was the same segments, but in the different order of 12, 4, and 8. The highest real-time OW and PW segments for the copilot at night were the same as those for the daytime PW ratings. The copilots commented during the post-mission debriefs that enroute segments had the greatest workload because of navigation and external communication responsibilities. As for the pilot, the FARP-SP segment had high workload, especially PW, because it included the engine failure.

The analysis of TLX subscale data revealed that, for the copilot, the Effort component of overall workload ratings was generally greater than that for the second and third most important components, Mental and Temporal Demands (103 and 75, respectively). The impact of the Effort component on overall workload ratings was particularly high for Segment 4 during both day and night missions. This latter finding probably reflects the additional effort required by the copilot during this particular mission segment. Here, in addition to the standard navigation tasks, the copilot had to assist the pilot by continuously monitoring aircraft speed and location, estimating time of arrival, and providing speed directions to the pilot to ensure that the fuel blivets were delivered on schedule.

DISCUSSION

The TAWL/TOSS Model

TAWL, it may be recalled, produces a timeline of workload at half-second intervals and determines the occurrence of "overload" for each of several separate channels or components of workload. The purpose of the current study was not to investigate the model's prediction of overload. Rather, the study focused on validating the underlying workload data base and the scenario generation rules developed for the TAWL/TOSS UH-60A model. Because the TAWL/TOSS model does not directly produce OW and PW values for each mission segment, a technique was developed to derive these values from the model output. The technique used to derive estimates of OW from the model output appears to be a reasonable method to predict real-time overall workload experiences. Indeed, high correlations were found between TAWL/ TOSS-derived OW scores and actual crew member real-time OW ratings (0.82 for 12 cases and 0.95 for 11 cases). These results lend confidence to the UH-60 workload data base and the scenario generation technique underlying the TAWL/TOSS model.

The correlations between TAWL/TOSS-derived PW scores and actual crew member PW ratings was significant but substantially lower (.62) than that found for the OW case. The inability of TAWL/ TOSS-derived PW scores to better discriminate differences in workload among mission segments may be attributed to the technique used to derive PW from the model output. For example, instead of selecting the maximum PW of any TAWL/TOSS half-second interval within a mission segment, it may be more meaningful to determine the maximum workload value of a longer time slice. This possibility was suggested by the conjecture that the crew estimates PW over a time interval longer than a half-second. In other words, the "psychological unit" is longer that one-half second, and it may be important for the TAWL/TOSS-derived PW to match this longer time unit. Furthermore, alternative schemes to determine PW in a single time-slice may employ the application of weights to each workload component before collapsing the data across components. Since the PW scale has not previously been used to assess workload, further research is necessary to determine the psychological nature of "peak workload" and thus the optimum PW computational method.

Workload and Operator Performance

No relationship was found between the independent rating of performance (IRP) and the crew member's real-time rating of workload. Specifically, the IRP was uniformly high. This result may be attributed to the scale employed by the observer to rate performance. The experimental performance scale was based on the rating system used by instructor pilots for evaluating students. In comparison to the performance of students, it is not surprising that the experimental crew members, all from the instructor pilot population, were given high performance ratings. That is, the pilots who participated in this study were experts themselves. They were highly proficient and capable of uniformly high levels of performance that are independent of workload.

Factor Validity

There were two methods utilized in the current UH-60A study to acquire validating information for the empirical workload measurement techniques. The first involved use of principal components analysis to determine if the scales were all measuring the same factor, in particular, a "workload" factor. Evidence for factor validity was found: the factor loadings of the four OWL techniques ranged from 0.8 to 0.9. The ordering of the factor validities of the four workload measures was TLX > OW > SWAT > MCH, similar to those found in earlier studies on diverse Army systems (e.g., Hill, Zaklad, Bittner, Byers, & Christ, 1988, and Byers, Bittner, Zaklad, & Christ, 1988). This result indicates that TLX has the highest factor validity (for the OWL "workload" factor) of the four measures used in the OWL Program studies.

The second validation method involved the collection of convergent data (Cook and Campbell, 1979). Specifically, OW, PW, and TLX numerical results were compared to the open-ended questionnaire data collected during the post-simulator flight interview. The interview results indicated a strong correspondence with the numerical reports concerning the distribution of workload across the missions and mission segments. A problem with this method is the fact that the same population was used to gather both the numerical workload scale ratings and the verbal interview responses. Due to time and resources constraints, we were unable to obtain verbal interviews concerning high and low workload

segments from an independent population of pilots. This problem may limit the convergent validity, but at the very least, illustrates the stability of the measurements within the same expert population.

Simulator and Real World Workload

The crew members participating in this study frequently commented that the workload experienced in the simulator differed from that experienced in an actual aircraft. In the simulator, there is no actual threat to life no matter what equipment failures, threat, or environmental conditions are encountered. Further, in another sense, performance in the actual aircraft is more critical than in the simulator because it can impact future career opportunities. Thus, motivation and possibly workload in the actual aircraft may be much higher than in the simulator.

On the other hand, the aviators also commented that in some cases workload in the simulator may be higher than in the aircraft for particular tasks. For example, the visual system of the simulator does not provide all the depth cues that would normally be provided in the aircraft. Such considerations indicate the need to follow-up with the crew members who participate in workload investigations to ensure that conclusions are properly drawn. As part of the OWL project, the results of this study were summarized and discussed with the group of pilots who participated in the study before this final report was written.

Real-time and Post-time Workload Ratings

Post-time (PT) ratings of OW and PW collected after a mission were found to be consistently lower than real-time (RT) ratings collected during the simulator flight. One possible explanation of this difference is that PT relies on memory which may be imperfect. This explanation is unlikely since an imperfect memory would produce errors in either direction and its net effect on mean workload ratings would be minimal. Further, if memory had decayed, PT ratings should have been closer to RT ratings for the segments nearer to the completion of the mission. The data do not reflect this. Workload ratings made during the post-mission session are consistently lower than those made real-time during all mission segments for pilots and in the majority of segments for the copilots.

Alternately, PT ratings may have been affected by the mere fact that the mission was completed. During the mission, two factors may have contributed to RT workload ratings: (a) the workload associated with the specific mission segment that was being rated; and (b) the workload associated with the uncertainty of anticipated future events during the mission. In this view, mission completion itself may have lowered the total subjective experience of workload. Thus, the PT measures may have reflected the workload associated with a set of specified task demands alone while the RT measures may have reflected all sources of workload. This speculation is supported by the fact that the difference between RT and PT ratings was greater for the night mission than for the day mission. The overall and general increase in difficulty associated with night missions may have led to greater real-time workload experiences during each segment of the flight as well as higher uncertainty of anticipated future events.

OW and PW Workload Ratings

The PW scale was a special measure devised specifically for this study. An issue associated with the introduction of a new scale is its sensitivity, or its ability to discriminate differences in task loading as well as to provide useful information that is otherwise unavailable. While the PW scale was shown to discriminate differences in workload, the ratings it produced were generally about 10 points higher than those produced by the OW scale. However, for Segment 12, the mean copilot PW rating was 19 points higher than the mean OW rating, indicating that a momentary peak had occurred during that segment that was qualitatively different from the peak workload that had occurred in any other segment. In fact, for both day and night missions, Segment 12 is given one of the highest ratings for PW but not for OW. These are reasonable findings considering the momentary nature of the simulated engine failure.

This finding does underscore the need to obtain measures of momentary workload as well as measures of workload "averaged" over an entire mission segment or task of interest. The sensitivity of PW to this difference alone, however, does not ensure its utility. Nevertheless, further research in the use of PW is warranted because the concept of peak workload is of critical importance. Even one brief instance of overload can lead to a mission failure in platform such as potentially unstable as the UH-60A.

Workload Scale Acceptance

The TLX scale received the highest overall favorable ratings by the aviators as the best descriptor of the workload that they experienced. The aviators preferred TLX because they could use it to rate workload on various subscales. The 100-point rating scale of the TLX scale was also preferred over the three-point scale of SWAT and the 10-point scale of the MCH technique. The OW, PW, and TLX scales were also considered to be the easiest scales to use. The MCH scale was rated as the hardest to use. Some crew members disliked the MCH scale because workload experience issues and major system design deficiencies were confounded. The aviators commented that they would have preferred that system deficiencies and workload issues be independently addressed. Some pilots felt that SWAT and MCH were too time consuming. The SWAT card sort required of the pilots prior to the experimental trials was also found to be objectionable. These results, like those for factor validity, were very similar to those found for other Army systems in the OWL Program (Byers et al., 1988 and Hill et al., 1988).

Pilot and Copilot Workload

In general, the pilots' workload was found to be higher for mission segments requiring pickup and landing zone operations, enroute flight while transporting an external load in a threat environment, and for the segment which included a simulated transient engine failure. For the copilot, workload was higher for enroute segments with threat and engine failure. Based on feedback from the crew members, these findings are reasonable and reflect workload that would be found in both the simulator and actual flight.

With the exception of the fuel blivet transport and the engine failure, the copilots' workload ratings were generally lower than the pilots' ratings. This latter finding may reflect the tasking of the crews during the experimental study. That is, prior to the simulator flight, the crew members were instructed not to share flight and navigation tasks during the mission as they normally would have during actual flight. These conditions were imposed upon the crew so that a clear comparison of pilot and copilot workload could be made across crews, missions, and mission segments. This would have been impossible if each crew used a different task allocation scheme. Thus, the finding that copilot workload was generally lower than pilot workload may not be found in actual flight during which the distribution of tasks (and workload) between the pilot and copilot may not only vary from that imposed during this study but could vary differentially as a function of mission segment.

CONCLUSION

The major conclusions drawn from this investigation are as follows.

1. The TAWL/TOSS model has shown a capability to reasonably track real-time empirical measures of workload. This finding indicates that TAWL/TOSS has substantial potential as an analytical technique that may be applied to predict workload early in the development cycle of a new system.

2. Empirical workload assessment techniques may be readily applied in an Army aviation setting with TLX and OW scales having the most favorable operator acceptance and the highest factor validity.

3. The PW scale may be a useful addition to the repertoire of workload rating scales following further research and validation.

REFERENCES

Bierbaum, C. R., Fulford, L. A., & Hamilton, D. B. (1989). Task analysis/workload (TAWL) user's guide (Technical Report ASI-690-323-89(a)). Ft. Rucker, AL: Anacapa Science, Inc.

Bierbaum, C.R., Szabo, S. M., & Aldrich, T. B. (1987). A comprehensive task analysis of the UH-60A mission with crew workload estimates and preliminary decision rules for developing a UH-60A workload prediction model (Technical Report MDA903-87-C-0523). Ft. Rucker, AL: Anacapa Science, Inc.

Bittner, A. C., Jr., Zaklad, A. L., Dick, A. O., Wherry, R. J., Jr., Herman, E. D., Bulger, J. P., Linton, P. M., Lysaght, R. J., & Dennison, T. W. (1987). Operator workload (OWL) assessment program for the Army: Validation and analysis plans for three systems (ATHS, Aquila, LOS-F-H). (TR 2075-3b). Willow Grove, PA: Analytics, Inc.

Byers, J. C., Bittner, A. C., Jr., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1988). Workload assessment of a remotely piloted vehicle (RPV) system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1145-1149). Santa Monica, CA: Human Factors Society.

Cook, T. D., & Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field setting. Chicago: Rand McNally.

Hart, S. G., & Staveland, L. E. (1987). Development of a NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. S. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: Elsevier.

Hill, S. G., Zaklad, A. L., Bittner, A. C., Jr., Byers, J. C., & Christ, R. E. (1988). Workload assessment of a mobile air defense missile system. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 1068-1072). Santa Monica, CA: Human Factors Society.

Hinkley, D. V. (1983). Jackknife methods. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.) Encyclopedia of statistical sciences: Vol. 4 (pp. 280-287). New York: Wiley.

Iavecchia, H. P., Linton, P. M. and Byers, J. C. Workload assessment during day and night missions in a UH-60 Blackhawk helicopter simulator. Proceedings of the Human Factors Society 33rd Annual Meeting (pp. 1481-1485). Santa Monica, CA: Human Factors Society.

Reid, G. B., Shingledecker, C. A., & Eggemeier, T. (1981). Application of conjoint measurement to workload scale development. Proceedings of the Human Factors Society 25th Annual Meeting (pp. 522-525). Santa Monica, CA: Human Factors Society.

Vidulich, M. A., & Tsang, P. S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proceedings of the Human Factors Society 31st Annual Meeting (pp. 1057-1061). Santa Monica, CA: Human Factors Society.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement application. Proceedings of the Human Factors Society 27th Annual Meeting (pp. 129-133). Santa Monica, CA: Human Factors Society.

DATA ATTACHMENT I-1

Real-time (RT) and Post-Mission (PT) Ratings for Overall and Peak Workload

Segment[a]		OW			PW		
No.	Name	RT	Est RT[b]	PT	RT	Est RT[b]	PT
		Pilot -- Day Mission					
1	SP-CP1	30.7	--	25.0	35.7	--	30.0
2	CP1-PZ	31.4	--	27.1	36.4	--	32.8
3	PZ Ops	39.3	53.0	36.4	50.0	62.4	44.3
4	PZ-LZ	45.0	38.1	40.7	54.3	70.2	51.4
5	LZ Ops	46.4	36.4	38.6	57.1	70.2	47.1
7	PZ Ops	38.6	--	39.3	50.0	--	47.8
8	PZ-Alt LZ	46.4	38.0	43.6	55.7	70.2	52.1
9	Alt LZ Ops	53.6	--	45.0	63.6	--	55.7
11	FARP Ops	32.8	29.8	25.0	40.7	70.2	30.7
12	FARP-SP	44.3	38.1	40.0	58.6	69.6	50.0
		Copilot -- Day Mission					
1	SP-CP1	24.3	--	22.1	29.3	--	30.0
2	CP1-PZ	27.1	--	29.3	32.8	--	36.4
3	PZ Ops	16.4	15.8	25.0	21.4	48.2	32.8
4	PZ-LZ	41.4	29.4	42.1	50.7	48.2	49.3
5	LZ Ops	29.3	22.7	25.7	37.1	48.2	33.6
7	PZ Ops	25.0	--	16.4	30.0	--	23.6
8	PZ-Alt LZ	38.6	35.0	34.4	46.4	50.9	43.6
9	Alt LZ Ops	31.4	--	21.4	39.3	--	27.1
11	FARP Ops	17.8	8.8	13.6	22.8	39.2	20.0
12	FARP-SP	39.3	30.2	34.3	57.8	48.2	52.8
		Pilot -- Night (NVG) Mission					
1	SP-CP1	43.6	--	36.4	52.8	--	44.3
2	CP1-PZ	42.1	--	40.0	55.0	--	47.8
3	PZ Ops	60.7	--	50.0	72.8	--	57.1
4	PZ-LZ	57.8	--	45.7	66.4	--	52.8
5	LZ Ops	66.4	--	52.1	76.4	--	60.7
7	PZ Ops	59.3	--	48.6	67.8	--	57.8
8	PZ-Alt LZ	50.7	--	48.6	60.0	--	57.8
9	Alt LZ Ops	57.8	-	50.7	65.7	--	58.6
11	FARP Ops	40.0	--	37.8	50.7	--	45.7
12	FARP-SP	57.1	--	54.3	70.0	--	62.8
		Copilot -- Night (NVG) Mission					
1	SP-CP1	41.4	--	36.4	50.0	--	44.3
2	CP1-PZ	41.2	--	38.6	50.0	--	46.4
3	PZ Ops	45.0	--	34.3	50.0	--	45.0
4	PZ-LZ	47.8	--	46.4	57.8	--	55.7
5	LZ Ops	39.3	--	36.4	52.1	--	42.8
7	PZ Ops	37.8	--	28.6	46.4	--	37.1
8	PZ-Alt LZ	49.3	--	46.4	60.0	--	58.6
9	Alt LZ Ops	40.7	--	40.0	47.1	--	51.4
11	FARP Ops	28.6	--	26.4	37.1	--	34.3
12	FARP-SP	52.1	--	45.0	65.0	--	63.6

[a] Segments 6 and 10 were not analyzed due to missing data.

[b] Est RT refers to IAWL/IOSS predictions of RT ratings; no such predictions were made for Segments 1 and 2 due to UH-60A simulator failures or for Segments 7 and 9 since they were identical to Segments 3 and 5, respectively.

DATA ATTACHMENT I-2

Task Load Index (TLX) Weighted Subscale Scores

Mission Segment	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration &
			Pilot -- Day Mission			
3	137	47	96	76	104	13
4	154	102	139	74	151	25
5	131	34	145	80	96	66
6	112	26	159	73	78	24
			Copilot -- Day Mission			
3	64	16	46	36	47	27
4	96	49	94	62	154	34
5	69	11	51	34	45	23
6	71	24	64	26	72	4
			Pilot -- Night (NVG) Mission			
3	193	69	156	96	140	30
4	174	88	121	66	130	30
5	160	38	209	70	106	99
6	146	37	200	74	111	41
			Copilot -- Night (NVG) Mission			
3	63	15	74	86	74	39
4	104	38	85	53	196	17
5	72	14	82	59	121	24
6	90	34	76	24	110	7